A STUDY OF BOOTSTRAP CONFIDENCE INTERVALS IN A COX MODEL

By
Deborah Burr

TECHNICAL REPORT No. 455
JULY 17, 1992

Prepared Under Contract
N00014-92-J-1254 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

A Study of Bootstrap Confidence Intervals in the Cox Model

Deborah Burr
Florida State University

June 1992

Abstract

We study bootstrap confidence intervals for three types of parameters in Cox's proportional hazards model: the regression parameter, the survival function at fixed time points, and the median survival time at fixed values of a covariate. Several types of bootstrap confidence intervals are studied, and the type of interval is determined by two factors. One factor is the method of drawing the bootstrap sample. We consider three such methods, which may be briefly described as follows: (1) Ordinary resampling from the empirical cumulative distribution function, (2) Resampling conditional on the covariates, and (3) Resampling conditional on the covariates and the censoring pattern. Another factor is the method of forming the confidence interval from a given sample; the methods considered are the percentile, hybrid, and bootstrap-t. We provide a theorem on the asymptotic validity of the third method of bootstrap resampling. All the methods of forming confidence intervals are compared to each other and to the standard asymptotic method via a Monte Carlo study. The data sets for this Monte Carlo study are simulated conditionally on the covariates and the censoring pattern, the situation appropriate for the third method of resampling. One conclusion drawn from the Monte Carlo study is that the asymptotic method is best for the regression parameter, but not for the survival function or the median survival time. Conclusions about the bootstrap methods include the surprising result that, overall, the second method of drawing the samples outperforms the third method. Also, there is an interaction effect between the two factors, method of drawing the sample and method of forming the interval, especially for estimation of the regression parameter. Finally, the bootstrap-t intervals are consistently outperformed by at least one of the two more rudimentary types of bootstrap interval.

Key words and phrases: ancillarity principle, bootstrap-t, hybrid interval, percentile interval, proportional hazards model

1 Introduction and Summary

The proportional hazards model of Cox (1972) specifies that the hazard rate for an individual with covariate vector x is

$$\lambda(t|x) = \lambda(t)\exp(\beta_0' x)$$

where β_0 is a vector of unknown regression coefficients and λ, the underlying baseline hazard rate, is an unknown and unspecified nonnegative function. Several parameters of common interest in the Cox model are the regression parameter β_0, the survival function $S(t|x)$ associated with $\lambda(t|x)$, and the p^{th} quantile of the distribution of the lifelength of an individual with covariate vector x, $\xi_p(x)$. The role of $\xi_{1/2}(x)$ is analogous to the role of the mean response, or the regression curve at the point x, in linear regression analysis.

For estimation of β_0 and $S(t|x)$, there exists a well-developed asymptotic theory which enables the construction of confidence intervals (Andersen and Gill (1982)). These confidence intervals have been found to work well in practice, and they are available in standard statistical computer packages, such as SAS and S. For estimation of $\xi_p(x)$, the results of Andersen and Gill (1982) must be applied in conjunction with a result on weak convergence of the quantile process to derive asymptotic theory (Dabrowska and Doksum (1987), Burr and Doss (1991)). The expression for the asymptotic standard error of the estimator involves the hazard rate function $\lambda(\xi_p(x))$, just as the formula for asymptotic standard error of the median of a simple random sample involves the underlying density function. Estimation of density functions or hazard rate functions is a complicated endeavor; Silverman (1986) discusses many methods which have been studied. Because the asymptotic theory for estimation of $\xi_p(x)$ is only recently available and requires estimation of the hazard rate, it is not widely applied.

This paper concerns the bootstrap method of forming confidence intervals in the Cox model. A principal reason for this study is that the bootstrap is an alternative to methods based on standard asymptotic theory. Before discussing further our motivation, it is useful to describe three distinct methods of bootstrapping in the Cox model.

The data and model may be described as follows. Associated with individual i are a covariate vector X_i, a lifelength Y_i, and a censoring time C_i. We do not observe Y_i directly, but rather we observe $T_i = \min(Y_i, C_i)$ and $\delta_i = I(Y_i \le C_i)$. Thus the data is (T_i, δ_i, X_i), $i = 1, \ldots, n$. The underlying baseline cumulative hazard function is $\Lambda(t) = \int_0^t \lambda(s)ds$. Then, the distribution function of the lifetime of an individual with covariate x is $F(t|x) = 1 - \prod_{u \le t}(1 - \Lambda(du))^{\exp(\beta_0' x)}$. (See Section 2.1.) If $\hat{\beta}$ and $\hat{\Lambda}$ are Cox's (1972) and Breslow's (1972, 1974) estimates of β_0 and Λ, respectively, then $F(t|x)$ may be estimated by $\widehat{F}(t|x) = 1 - \prod_{u \le t}(1 - \hat{\Lambda}(du))^{\exp(\hat{\beta}' x)}$. Assume that the C_i's are iid $\sim G$, and let $\widehat{G}$ be the Kaplan-Meier estimate of G based on the data (T_i, δ_i, X_i), $i = 1, \ldots, n$.

Consider the following two methods of bootstrapping:

Method 1: Resample the triples (T_i, δ_i, X_i), $i = 1, \ldots, n$.

Method 2: Generate $Y_i^* \sim \widehat{F}(t|X_i)$ and $C_i^* \sim \widehat{G}$, $i = 1, \ldots, n$, all variables independent. Form $T_i^* = \min(Y_i^*, C_i^*)$ and $\delta_i^* = I(Y_i^* \le C_i^*)$. The resampled data is then (T_i^*, δ_i^*, X_i), $i = 1, \ldots, n$.

Efron and Tibshirani (1986) discuss Method 1. In an interesting but unpublished technical report which provided impetus for the present work, Hjort (1985) proposes

Method 2 and develops some asymptotic theory for it. Suppose that we wish to estimate the variability of some estimate, such as $\hat{\xi}_p(x)$. Method 1 is appropriate for estin.ating the unconditional variance of $\hat{\xi}_p(x)$, i.e. averaging over the marginal distribution of the covariates and of the censoring variables. Method 2 is appropriate for estimating the conditional variance of $\hat{\xi}_p(x)$ given X, where $X = (X_1, \ldots, X_n)$. If the distribution of the X_i's does not depend on the unknown parameters S and β_0 then the usual ancillarity arguments point to the conditional variance as the "right" quantity to estimate. This situation is closely connected to bootstrapping in linear regression models, where one can bootstrap by resampling from the pairs (responses, covariates), or one can bootstrap by resampling from the residuals; see Freedman (1981). Many of the comments in the discussion paper Wu (1986) are relevant here.

The ancillarity principle can be carried further in the presence of censoring, where the *censoring pattern* is an ancillary statistic. The distribution of the C_i's does not depend on the unknown parameters S and β_0; therefore, if we knew the C_i's we would condition on them. However, we don't see the C_i's exactly: If $\delta_i = 0$ we see the exact value of C_i, but if $\delta_i = 1$ we know only that $C_i > T_i$. Denote this information on the C's by $\mathcal{C}$. Then, what we want to estimate is $\text{Var}(\hat{\xi}_p(x)|X, \mathcal{C})$. This idea leads to Method 3 of bootstrapping.
Method 3: Generate $Y_i^* \sim \hat{F}(t|X_i)$. If $\delta_i = 0$, let $C_i^* = T_i$; if $\delta_i = 1$, generate C_i from the Kaplan-Meier estimate of G, conditional on $C_i > T_i$, that is, from the distribution $(\hat{G}(t) - \hat{G}(T_i))/(1 - \hat{G}(T_i))$.

In deciding to study Method 3 of taking bootstrap samples, we were motivated by consideration of a situation where this method appears to provide a large improvement over Method 2. Suppose β_0 is positive, so that large covariate values are associated with large risk. Then a data point with a large covariate value, for which the lifetime is large and uncensored, would be an influential point. The estimator $\hat{\beta}$ is pulled heavily downward by such an influential point. For estimation of the bias of $\hat{\beta}$ conditional on the covariate/censoring pattern, it appears intuitively true that Bootstrap Method 3 is better than Method 2. The large lifetime will usually be censored in Method 2 bootstrap samples, and in effect these bootstrap samples will be very similar to those arising if the influential point did not exist. In contrast, Method 3 forces the large lifetime to remain uncensored, and so this problem does not arise.

In the preliminary study that motivated our consideration of Method 3, when we compared the bootstrap estimates of bias of $\hat{\beta}$ obtained from the three methods of taking bootstrap samples, Method 3 had the lowest m.a.d. (median absolute deviation from the true value) of the three methods. Details of an expanded version of this study are given in Section 3.

We can now discuss more fully the motivations for the present work. First, although asymptotic methods already exist for the main interesting parameters, in the case of $\xi_p(x)$ they are difficult to apply. Burr and Doss (1991) apply the asymptotic theory to formation of confidence bands for $\xi_p(x)$ and show good performance of the bands in some Monte Carlo studies. They use a kernel function method to estimate $\lambda(\xi_p(x))$, for which the bin width was carefully chosen to be suitable for the particular simulation study. They caution that if it is not possible to exert such care in selecting the bin width in practice, then the performance of the bands may be adversely affected. In addition, even when the asymptotics are easily implemented, the bootstrap may outperform the asymptotic methods. Singh (1981) and Abramovitch and Singh (1985) give situations where the sampling variability of an estimator is more accurately estimated using bootstrap procedures than using standard asymptotic methods.

A second reason for this investigation on bootstrap confidence intervals in the Cox model is that "the bootstrap" method of forming confidence intervals is not in fact uniquely defined. There are two aspects to this problem: First, it is often true that when dealing with complex data structures there are several ways to draw the bootstrap samples; second, many methods for forming bootstrap confidence intervals from a given sample have been proposed. A major thrust of the present work is to study the effects of method of taking the sample and method of forming the interval, on performance of the bootstrap in interval estimation of the parameters of most interest in the Cox model.

So many types of bootstrap confidence intervals have been proposed that the variety available can be overwhelming. Types which are frequently studied in the literature include the percentile, hybrid, bootstrap-t, and Efron's BC_a, all of which are discussed by Martin (1990). Here we study only three kinds of bootstrap confidence intervals in order to keep the work focused, and to simplify the conclusions drawn. We study the percentile, hybrid, and bootstrap-t intervals. These intervals are described in Section 2.2 below.

The percentile method is considered because even though the theoretical justification for this method is the weakest (Hall (1988)), these intervals are the simplest to use and explain, and are the most frequently used in practice. We study the hybrid method because this is precisely the method which is justified by asymptotic results for the bootstrap in complicated models, such as the Cox model. That is, let θ represent a parameter to be estimated, $\hat{\theta}$ an estimate of it, and $\hat{\theta}^*$ the estimate computed from a bootstrap sample. Suppose we know that $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ has almost surely the same limiting distribution as $\sqrt{n}(\hat{\theta} - \theta)$. Then we would want to use the distribution of $(\hat{\theta}^* - \hat{\theta})$ to approximate that of $(\hat{\theta} - \theta)$; if we do this to form confidence intervals for θ, then the intervals obtained are the hybrid ones. The bootstrap-t and the BC_a intervals are comparable in that both have been demonstrated theoretically to be "second-order correct" for one-sided intervals in some relatively simple situations; see Hall (1988). The bootstrap-t stood out as a star performer in recent empirical work of Owen (1988), which dealt with nonparametric interval estimation of the mean from a random sample. In addition, the bootstrap-t is usually more automatic to apply than the BC_a method. For these reasons, we study the bootstrap-t rather than the BC_a method in this work. However, we must also note here a problem with the bootstrap-t which was especially apparent in our work on applying it to forming confidence intervals for $\xi_p(x)$. The recent work which has shown good behavior of the bootstrap-t has dealt with simple cases, such as estimating the mean of a random sample. It is well-known that in more complicated situations, stability of the estimate of scale needed for the bootstrap-t is crucial to its good performance. It may be very difficult to estimate this scale parameter, and in fact, the bootstrap-t method is not uniquely defined since many different scale parameters, and different estimators of these parameters, could be tried. In this work, we have taken the approach of building on known asymptotic results in order to lessen the computational labor of the bootstrap-t; that is, we estimate the standard error of our estimators by formulas derived from asymptotic theory. This approach did not work for $\xi_{1/2}(x)$. See Section 2.2 for a discussion of the bootstrap-t for estimation of $\xi_p(x)$, and for detailed descriptions of the three methods we study.

Asymptotic results for estimators in the Cox model are based on martingale central-limit theorems. It is an unfortunate feature of this martingale theory that it is capable of producing only first-order results. That is, there are no tools available which are analogous to the Edgeworth expansions that yield the currently available theoretical comparisons of the various types of intervals. When faced with the impossibility of obtaining a theoretical comparison of the various bootstrap methods for forming confidence intervals using the existing theory, comparisons must be done by Monte Carlo studies. Here, the Monte Carlo studies are carried out in the situation appropriate for Method 3 of bootstrapping, that is, conditional on the covariates and the censoring pattern; this is explained in more detail in Section 3.1.

There is an enormous number of possible Monte Carlo studies that could be carried out, particularly since we condition on the covariates and the censoring pattern. We have carried out simulations for a wide range of the factors involved. In choosing which studies to present here, the decision was complicated by inconsistencies in the results among the different situations we studied. In particular, there were marked changes in the relative performances of the different methods for low and high percent censoring. We present in detail in this paper the results for a situation with high percent censoring (55%), over several sample sizes and several covariate/censoring patterns. We selected the high percent censoring case because it is our experience that many applications of the Cox model involve high percent censoring due to a fixed endpoint of the clinical trial. We also report results of a single study with an influential point of the sort described above in our motivation for Method 3 of bootstrapping.

Before carrying out the Monte Carlo studies, we had anticipated that confidence intervals would improve with increasing sophistication of the method of drawing bootstrap samples and method of forming the intervals. However, this was not the case. Even in the restricted set of studies we report on here, the results are not simple to describe; there is no single "winner" among the bootstrap methods. To begin with, different conclusions are drawn for each type of parameter studied, β_0, $S(\cdot)$, and $\xi_p(\cdot)$. For forming confidence intervals for β_0, the asymptotic method is consistently better than, or at least as good as, all the bootstrap methods considered. This is not true for the confidence intervals for $S(\cdot)$ or $\xi_p(\cdot)$. Nevertheless, we venture to make the following general statements here. One surprising result is that Method 2 overall outperforms Method 3, except in the study with the influential point. Also, there is an interaction effect between the two factors, method of drawing the sample and method of forming the interval, especially for estimation of β_0. Finally, the bootstrap-t intervals are consistently outperformed by at least one of the two more rudimentary bootstrap methods.

In Section 5 we deal with asymptotics, and we consider the simpler case where β_0 is known to be zero. This corresponds to the familiar setup of the Kaplan-Meier estimator. In the case of the Kaplan-Meier estimator, Methods 1 and 2 of bootstrapping are identical (Efron (1981)). It has been shown for Method 1 (or Method 2) that the Kaplan-Meier estimator computed from the bootstrap sample, when standardized, converges weakly to the same Gaussian process to which the standardized Kaplan-Meier estimator itself converges (Akritas (1986); Lo and Singh (1986)). We show the same result for Method 3. Thus Methods 1 and 3 may be regarded as "asymptotically equivalent."

The rest of the paper is organized as follows: Section 2 gives a detailed description of the notation and algorithms needed for estimation of the Cox model parameters and for description of the bootstrap confidence intervals. Section 3 describes the set-up for the Monte Carlo studies and reports results of the studies. Section 4 gives a summary and states conclusions. The theoretical result is stated and proved in Section 5. The figures and chart for Section 3 are contained in the appendix.

2 Notation and Algorithms

2.1 Estimation of the Cox Model Parameters

Here we describe the estimators of β_0, $S(t|x)$, and $\xi_p(x)$, and the estimators of the standard errors of these estimators. Several estimators are possible for each of these parameters. The particular estimator used has an impact on the bootstrap procedure, particularly because of the many ties in the bootstrap samples; therefore it is necessary to include many details in this description, which will be given in counting process notation, following closely that of Andersen and Gill (1982) (henceforth AG).

In the counting process formulation of the likelihood, we use X_i, the q-dimensional vector of covariates, but rather than using T_i and δ_i directly we instead define the counting processes

$$N_i(t) = I(T_i \leq t,\ \delta_i = 1) \quad \text{for } t \geq 0$$

and the processes

$$J_i(t) = I(T_i \geq t) \quad \text{for } t \geq 0.$$

In this notation, conditional on $X_i = x_i$, $i = 1, \ldots, n$, the partial likelihood of β_0 at time τ is

$$L(\beta, \tau) = \prod_{u \in [0,\tau]} \prod_{i=1}^{n} \left(\frac{J_i(u) \exp(\beta' x_i)}{\sum_{j=1}^{n} J_j(u) \exp(\beta' x_i)} \right)^{dN_i(u)}.$$

The maximum partial likelihood cstimator of β_0 at time τ is the value $\hat{\beta} = \hat{\beta}(\tau)$ of β that maximizes $L(\beta, \tau)$. In practice, of course, one uses the value of β that maximizes the partial likelihood at time ∞; see the discussion in Section 4 of AG. In the case of ties, we use the usual Peto approximation (Peto (1972)). Other solutions are possible, but they are more computationally intensive; we considered such solutions impractical because the number of ties in the bootstrap samples can be quite large.

Next we must specify the estimators of $S(t|x)$ and $\xi_p(x)$, and for this it is first necessary to give an estimator of Λ. The "Nelson-Aalen" estimator of Λ is

$$\hat{\Lambda}(t) = \int_0^t \Big(\sum_{j=1}^{n} J_j(s) \exp(\hat{\beta}' x_j) \Big)^{-1} d\Big(\sum_{i=1}^{n} N_i(s) \Big).$$

The estimator $\hat{\Lambda}(t)$ increases only by jumps, which occur at the uncensored deaths. Breslow's (1972, 1974) estimator of $\Lambda(t)$ is the continuous estimator obtained from the Nelson-Aalen estimator by linear interpolation between observed failure times. There are two general approaches to estimation of $S(t|x)$ available in practice. The Tsiatis/Link/Breslow approach takes an estimator of the form

$$\hat{S}(t|x) = \exp(-\tilde{\Lambda}(t|x)),$$

for some estimator $\tilde{\Lambda}$ of Λ. Tsiatis (1983) uses this form for estimation of $\hat{S}(t)$ with the Nelson-Aalen estimator of $\Lambda(t)$; Link (1984) takes the same form but uses the Breslow estimator of $\Lambda(t)$. We note that the relationship

$$S(t) = \exp(-\Lambda(t))$$

is only valid for continuous T. The other approach is to use the product integral; Kalbfleisch and Prentice (1980, p. 86) provide a nonparametric MLE for $S(t|x)$ using this approach. We use the product integral, which is described as follows. Note that for an *arbitrary* cumulative hazard function H (which may contain discrete or continuous components or both), the survival function corresponding to H is the product integral

$$S(t) = \prod_{s \leq t}(1 - H(ds)). \tag{2.1}$$

See Gill and Johansen (1990) or Kalbfleisch and Prentice (1980, sec. 1.2.3). Then, given an estimator $\tilde{H}(t)$ of H, the product-integral estimator of $S(t)$ is

$$\hat{S}(t) = \prod_{s \leq t}(1 - \tilde{H}(ds)). \tag{2.2}$$

To specify the estimator of $S(t)$ it remains to give our estimator of the cumulative hazard function. We use the Nelson-Aalen estimator of $\Lambda(t)$, which is a step-function estimator. (We later do a linear smooth of $\hat{S}(t|x)$; see Remark 3 below.) In this paper we study the baseline survival function; however, we also need to decide on an estimator of $S(t|x)$ for the purpose of defining an estimator of $\xi_p(x)$. Taking care to specify the model appropriately for an arbitrary distribution, e.g. such as a discrete bootstrap distribution, we use the following relationship between $\Lambda(t|x)$ and $\Lambda(t)$

$$1 - \Lambda(dt|x) = (1 - \Lambda(dt))^{\exp(\beta_0' x)}. \tag{2.3}$$

See Kalbfleisch and Prentice (1980, sec. 2.4.2 & 4.6.1). The above relationship combined with 2.2 yields the following estimator of $S(t|x)$

$$\hat{S}(t|x) = \prod_{s \leq t}(1 - \hat{\Lambda}(ds))^{\exp(\hat{\beta}' x)}. \tag{2.4}$$

A more detailed explanation for this estimator appears in Burr and Doss (1991).

Remarks on Computational Details

1 As is commonly done in fitting the Cox model, we assume covariates have been centered at their mean; so, we take the baseline hazard function and baseline survival function to be at the mean covariate values. That is, to be fully precise, in Equation (2.4), the Nelson-Aalen estimate $\hat{\Lambda}(t)$ of $\Lambda(t)$ is computed at mean-centered covariates, and the

covariate vector x in the exponent must be mean-centered. To keep notation as simple as possible, from now on we assume covariates have already been corrected for the mean. In this paper we study the estimator of the baseline survival function, $\hat{S}(t|\overline{x})$, which we refer to from now on as $\hat{S}(t)$.

2 If any factor $(1 - \hat{\Lambda}(ds))$ is less than zero, then $\hat{S}(t|x)$ is taken to be zero.

3 The function $\hat{S}(t|x)$ is a step function, with jumps at the observed uncensored survival times. We actually use here a smoothed version of $\hat{S}(t|x)$, $\hat{S}^c(t|x)$, which may be described as follows: Say that the number of unique uncensored survival times, plus the last censored observation if it is the last observation overall, is n_u, with ordered values denoted $T_{u(1)}, \ldots, T_{u(n_u)}$. Also, set $T_{u(0)} = 0$. For $0 \leq t \leq T_{u(1)}/2$, $\hat{S}^c(t|x) = 1$. Let $W_{u(i)} = \hat{S}(t_{u(i-1)}|x) - \hat{S}(t_{u(i)}|x)$ be the weight assigned to the i^{th} uncensored observation by $\hat{S}(t|x)$ as given in Equation (2.4) and Remarks 1 and 2 above. Then "smear" half of $W_{u(i)}$ left, the other half right, where the smearing is done by linear interpolation from $T_{u(i)}$ to halfway between $T_{u(i)}$ and the adjacent uncensored observation. If $n_u = 0$ the estimator is undefined; these cases were skipped in the bootstrap resampling. Beyond the last observation $T_{u(n_u)}$, the linear extrapolation uses the slope of the previous segment, until the point t_0 is reached for which $\hat{S}^c(t_0|x) = 0$; then for $t > t_0$, $\hat{S}^c(t|x) = 0$. Note that as a result of this definition, the survival function does eventually decrease to zero, even if the last observation is censored.

To estimate $\xi_p(x)$ we note that the p^{th} quantile of $1 - S$ is $(1 - S)^{-1}(p)$. Here and throughout the paper, for an arbitrary increasing function f, f^{-1} denotes the right continuous inverse of f defined by $f^{-1}(t) = \sup\{s : f(s) \leq t\}$. For the case of the survival function given by Equations (2.1) and (2.3), this gives

$$\xi_p(x) = \sup\Big\{s : 1 - \prod_{u \leq s}(1 - \Lambda(du))^{\exp(\beta_0' x)} \leq p\Big\}. \tag{2.5}$$

Substituting $\hat{\Lambda}$ for Λ and $\hat{\beta}$ for β_0, we obtain the estimate

$$\hat{\xi}_p(x) = \sup\Big\{s : 1 - \prod_{u \leq s}(1 - \hat{\Lambda}(du))^{\exp(\hat{\beta}' x)} \leq p\Big\}. \tag{2.6}$$

In fact, we use a continuous version $\hat{\xi}_p^c(x)$ of $\hat{\xi}_p(x)$ based on the continuous version of $\hat{S}(t|x)$ described in Remark 3 above. That is, $\hat{\xi}_p^c(x)$ can be obtained by solving for t in the equation $\hat{S}^c(t|x) = 1 - p$. From now on, for the sake of brevity of notation, the superscript c is omitted in $\hat{S}^c(t|x)$ and $\hat{\xi}_p^c(x)$.

Further notation is needed in order to give the formulas for the standard errors of $\hat{\beta}$, $\hat{S}(t)$, and $\hat{\xi}_p(x)$. The notation below follows closely that of AG. For a q-vector $w = (w_1, \ldots, w_q)$, $w^{\bigotimes 2}$ denotes the $q \times q$ matrix whose $(i,j)^{\text{th}}$ entry is $w_i w_j$. Define

$$\begin{aligned} S^{(0)}(\beta,t) &= \frac{1}{n}\sum_{l=1}^{n} J_l(t)\exp(\beta' x_l), \\ S^{(1)}(\beta,t) &= \frac{1}{n}\sum_{l=1}^{n} x_l J_l(t)\exp(\beta' x_l), \\ S^{(2)}(\beta,t) &= \frac{1}{n}\sum_{l=1}^{n} x_l^{\otimes 2} J_l(t)\exp(\beta' x_l), \\ E(\beta,t) &= \frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)}, \end{aligned} \qquad (2.7)$$

and

$$V(\beta,t) = \frac{S^{(2)}(\beta,t)}{S^{(0)}(\beta,t)} - (E(\beta,t))^{\otimes 2}.$$

It is shown in AG that a consistent estimator of the asymptotic covariance matrix of $\sqrt{n}(\hat\beta - \beta_0)$ is given by

$$\hat\Sigma^{-1}_{\sqrt{n}\hat\beta} = \left\{\int_0^{\tau} V(\hat\beta,t)S^{(0)}(\hat\beta,t)d\hat\Lambda(t)\right\}^{-1}. \qquad (2.8)$$

Next we discuss estimation of the standard error of $\hat S(t)$. The results of AG may be used to easily show that the asymptotic variance of $\sqrt{n}(\hat\Lambda(t) - \Lambda(t))$ may be consistently estimated by

$$\hat\sigma^2_{\sqrt{n}\hat\Lambda(t)} = \hat a(t) + \hat b'(t)\hat\Sigma^{-1}_{\sqrt{n}\hat\beta}\hat b(t), \qquad (2.9)$$

where

$$\begin{aligned} \hat a(t) &= \int_0^t (S^{(0)}(\hat\beta,u))^{-1} d\hat\Lambda(u), \\ \hat b(t) &= \int_0^t E(\hat\beta,u)d\hat\Lambda(u). \end{aligned} \qquad (2.10)$$

Since $\hat S(t) = \prod_{s\leq t}(1 - \hat\Lambda(ds))$, we can apply a functional version of the δ-method to obtain the asymptotic variance of $\sqrt{n}(\hat S(t) - S(t))$. See Gill and Johansen (1990). On a less technical level, we can write $\hat S(t) \doteq \exp(-\hat\Lambda(t))$ and apply the standard δ-method. The two answers obtained are identical, and the variance so obtained can be estimated consistently by

$$\hat\sigma^2_{\sqrt{n}\hat S(t)} = \hat\sigma^2_{\sqrt{n}\hat\Lambda(t)}\exp(-2\hat\Lambda(t)). \qquad (2.11)$$

Dabrowska and Doksum (1987) and Burr and Doss (1991) derive the asymptotic variance of $\hat\xi_p(x)$, which depends upon the baseline hazard rate $\lambda(t)$. To describe our estimator of the asymptotic variance of $\hat\xi_p(x)$, we must define our estimator of the baseline hazard rate $\lambda(t)$. Possibilities include methods based on splines (Whittemore and Keller (1986)) and those based on kernel smoothers. Kernel smoothers are computationally convenient, and in addition their asymptotic properties in the present context have already been studied by Ramlau-Hansen (1983). To describe them, let R be a function of bounded variation with support on $[-1,1]$, and whose integral is 1, and let $\{b_n\}$ be a sequence of positive constants such that as $n \to \infty$, we have $b_n \to 0$ and $nb_n^2 \to \infty$. Define the kernel estimate of $\lambda(\cdot)$ by

$$\hat{\lambda}(t) = \frac{1}{b_n}\int_0^\infty R\Big(\frac{t-s}{b_n}\Big)d\hat{\Lambda}(s). \tag{2.12}$$

The specific choices of R and $\{b_n\}$ are discussed, in the context of ordinary density estimation, in Silverman (1986, pp. 40–72). Details of our algorithms are provided in the Remarks below. Having specified a choice of R and $\{b_n\}$ we may define an estimate of $\sigma^2(\sqrt{n}\hat{\xi}_p(x))$. Using the notation $\pi = \log(1-p)$, we have the estimate

$$\hat{\sigma}^2_{\sqrt{n}\hat{\xi}_p(x)} = \frac{\hat{a}(\hat{\xi}_p(x))}{(\hat{\lambda}(\hat{\xi}_p(x)))^2} + \Big(\frac{\hat{b}(\hat{\xi}_p(x)) + \pi x \exp(-\hat{\beta}'x)}{\hat{\lambda}(\hat{\xi}_p(x))}\Big)\hat{\Sigma}^{-1}\Big(\frac{\hat{b}(\hat{\xi}_p(x)) + \pi x \exp(-\hat{\beta}'x)}{\hat{\lambda}(\hat{\xi}_p(x))}\Big). \tag{2.13}$$

Remarks on the Kernel Estimator of the Hazard Rate

1 In this paper we use the biweight kernel

$$R(t) = \frac{15}{16}(1-t^2)^2 \qquad |t| < 1 \tag{2.14}$$

2 In choosing the window widths b_n we have followed Ramlau-Hansen (1983) in allowing varying window widths over the range of survival times. These would generally be increasing as t increases to cope with the scarcity of data for larger t. In particular, we allowed four different intervals of the time axis with different window widths in each. The window widths were taken to be inversely proportional to $n^{1/3}$, and the choice of constant of proportionality was made subjectively, using one or two simulated data sets for each simulation situation.

2.2 Description of Confidence Intervals

Four methods are used to form approximate confidence intervals for the parameters β_0, $S(t)$, and $\xi_p(x)$. Here we first give brief descriptions of these methods, followed by further discussion of the bootstrap-t. We also describe a method of refining each of the types of bootstrap intervals, called the iterated bootstrap, which we do not study in this work, but which we refer to in discussions of our results.

Denote the parameter being estimated by θ, its estimator by $\hat{\theta}$ and the estimate of standard error of $\hat{\theta}$ by $\hat{\sigma}$. In forming the confidence intervals described below, we have used the estimators $\hat{\theta}$ and $\hat{\sigma}$ defined in Section 2.1. The nominal approximate coverage probability of the intervals is $100(1-2\alpha)$. The $(1-\alpha)^{\text{th}}$ quantile of the standard normal distribution is denoted $z^{(\alpha)}$. The standard method, based on asymptotic theory for the Cox model, gives the interval

$$(\hat{\theta} \pm \hat{\sigma} z^{(\alpha)}). \tag{2.15}$$

Three bootstrap methods are studied. We refer to an estimate computed from a bootstrap sample as $\hat{\theta}^*$, and we let K denote the bootstrap distribution of $\hat{\theta}^*$. The percentile interval is

$$(K^{-1}(\alpha), K^{-1}(1-\alpha)). \tag{2.16}$$

In the hybrid method, the bootstrap distribution of $\hat{\theta}^* - \hat{\theta}$ is used to get approximate quantiles of the distribution of $\hat{\theta} - \theta$. This leads to the interval

$$(2\hat{\theta} - K^{-1}(1-\alpha), 2\hat{\theta} - K^{-1}(\alpha)). \tag{2.17}$$

See Efron (1990, p. 14). For the bootstrap-t, consider the bootstrap distribution of $T^* = (\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$, where $\hat{\sigma}^*$ is the estimate $\hat{\sigma}$ computed from a bootstrap sample. Call this bootstrap distribution K_t. The bootstrap-t interval is

$$(\hat{\theta} - {K_t}^{-1}(1-\alpha)\hat{\sigma}, \hat{\theta} - {K_t}^{-1}(\alpha)\hat{\sigma}). \tag{2.18}$$

In the Monte Carlo studies reported in Section 3, the four types of confidence intervals described above are studied for all the parameters considered, except that bootstrap-t intervals are not studied for $\xi_{1/2}(\cdot)$. Because estimates of the standard error of $\xi_{1/2}(\cdot)$ are inherently unstable, especially in situations with high censoring, the bootstrap-t using the asymptotic formula for $\hat{\sigma}$ did very poorly. One could attempt use of the *bootstrap* estimate of standard deviation of $\hat{\xi}_p(x)$ in the denominator of the bootstrap-t. That is, in Equation 2.18 above, use the bootstrap estimate of σ rather than the asymptotic formula for $\hat{\sigma}$. Note that then, in order to compute the bootstrap statistic T^*, the denominator $\hat{\sigma}^*$ is obtained through a second layer of bootstrapping, and the amount of time that this takes on a single sample is so large that Monte Carlo studies of this method are impossible on workstations commonly available today. Also, even with the bootstrap estimate of standard deviation, the bootstrap-t may do poorly. Doss and Gill (1991), in an example where they are estimating the quantiles of the survival function in the random censorship model of survival analysis, had difficulties with unstable bootstrap-t intervals when they attempted to use the bootstrap estimate of standard deviation. On a theoretical level, the performance of bootstrap confidence intervals for quantiles takes on a different character from their performance for quantities such as the mean: The difference between nominal and actual coverage is $O(n^{-1/2})$ not $O(n^{-1})$; that is, these intervals are only first-order accurate, not second-order. See the related work of Hall and Martin (1988). We make one final comment here, in which is indicated some hope for the bootstrap-t in this situation: Doss and Gill (1991) try other measures of scale for the denominator of the bootstrap-t in their example, and they settle on a particular interquantile range which led to stable bootstrap-t intervals. However, we have not attempted to do this here.

In this context, it is important to mention some recent research on the iterated bootstrap method of refining confidence intervals which was discussed by Beran (1987). Descriptions of the use of the method for coverage correction of confidence intervals may be found in Martin (1990) and DiCiccio, Martin and Young (1990). Here we give a brief description of the idea behind this method. Suppose we want to form a 90% percentile confidence interval. The idea is to estimate the actual coverage probability of the percentile intervals for several nominal levels, and then use the percentile interval which has estimated coverage probability exactly equal to the desired level of 90%. The estimation of coverage probabilities is done through a second layer of bootstrapping; that is, from each bootstrap sample used in forming the original confidence interval, many bootstrap samples are drawn and the percentile interval formed. So, this method is extremely computationally intensive. The recent work of DiCiccio, Martin and Young (1990) is on an analytic method to replace the Monte Carlo simulation in the inner layer of bootstrap sampling, for the problem of constructing confidence intervals for a parameter θ that is expressible as a smooth function of vector means. Their work does not apply to bootstrapping in the Cox model. We did not attempt the iterated bootstrap in our work, but we refer to it in discussion of the results of the Monte Carlo studies.

3 The Monte Carlo Studies

First, we mention an important aspect of our discussion of the Monte Carlo results. We compare the various methods in terms of coverage probability and average or median length. A method with higher coverage probability and smaller average length than a second method is certainly better than the second method. However, we often find that the method with higher coverage probability also has larger average length, and in this case it may be that the iterated bootstrap would produce a good interval. Or, one could devise a way to compare the methods by adjusting all lengths to be equal and then comparing coverage probabilities, as in Owen (1988). We have not done this here; rather, we discuss the results simply through direct comparisons of coverage probability and average length, which generally leads us to call some method "best" for a particular situation if its coverage probability is the closest to the nominal level and its length is not exorbitant relative to the other methods. In our choices of the "winners," we admit to some conservative prejudice in favor of accuracy of coverage probability over shortness in length.

In Section 3.1, we describe in a general way how we simulate the data sets conditionally on the covariate/censoring pattern. In Section 3.2, we list the factors determining the simulations and state the levels of these factors which were included in the main computer experiment which we report on here. In Section 3.3, we summarize the results of this main study. Finally, in Section 3.4, we describe the study with an influential point and state briefly the main results from it.

3.1 General Description of a Simulated Dataset

Assume we have decided upon the sample size n, the regression parameter β_0, the covariate distribution F_X, the lifetime distribution F, and the censoring distribution G. The data sets in our simulation studies were generated *to be compatible with a fixed covariate/censoring pattern* according to the following steps.

1 Get one set of data (X_i, T_i, δ_i), $i = 1, \ldots, n$ as follows: First, generate the covariate values $X_i \sim F_X$, the survival times $Y_i \sim F(t|X_i) = 1 - (1 - F(t))^{\exp(\beta_0' X_i)}$, and the censoring times $C_i \sim G$; next, form the data points (X_i, T_i, δ_i) where $T_i = \min(Y_i, C_i)$ and $\delta_i = I(Y_i \leq C_i)$.

2 Generate one data set compatible with the covariate/censoring pattern $(X, \mathcal{C})$ obtained in Step 1, as follows. For $i = 1, \ldots, n$,

a. Let $X_i' = X_i$;

b. Generate Y_i' as in Step 1.

c. Generate C_i' conditionally on C, as follows. If $\delta_i = 0$, then $C_i' = T_i$. If $\delta_i = 1$, then generate C_i' from the conditional distribution of C given that $C > T_i$. For example, if G is the Exponential(1) distribution, then using the memoryless property of the exponential distribution, we let $C_i = T_i + rexp()$, where $rexp()$ denotes a random draw from the Exponential(1) distribution.

d. Form $T_i' = \min(Y_i', C_i')$, $\delta_i' = I(Y_i' \leq C_i')$.

3 Repeat Step 2 many times. (Each time get confidence intervals for the parameters of interest by the several methods and record relevant information such as whether the intervals contain the true value, and the length of the intervals.)

3.2 Factor Levels Included in the Experiment

From the above description of a simulated data set, it is clear that the factors affecting a simulation are:

1 F_X, the covariate distribution

2 F, the lifetime distribution

3 The Cox regression parameter β_0

4 The form of the censoring distribution G

5 The average percent censoring

6 The sample size n

7 The particular covariate pattern

8 The particular censoring pattern

In the study we describe here, we use the Uniform(0,1) distribution for the covariates. Also, we use the standard Exponential distribution for F, we take $\beta_0 = 2$, and we use the Uniform distribution for G. We take the mean of the censoring distribution to be .25 for approximately 55% average amount of censoring. The sample sizes considered are $n = 30, 40, 50, 60, 70, 80, 90, 100$. For the smaller sample sizes of $n = 30$, $n = 40$, and $n = 50$, there are three distinct covariate/censoring patterns; for the larger sample sizes of $n = 60$ and up, there are two distinct covariate/censoring patterns.

We report 90% confidence intervals for the parameters β_0, $S(.106)$, $S(.255)$, $\xi_{1/2}(.5)$, and $\xi_{1/2}(.939)$. The values of t at which the function $S(\cdot)$ is studied, $t_1 = .106$ and $t_2 = .255$, are the .25 and .50 quantiles of the distribution of $S(t|x = .5)$. The values of x at which the function $\xi_{1/2}(\cdot)$ is studied, $x_1 = .5$ and $x_2 = .939$, are such that $\xi_{1/2}(x_1) = t_2$ and $\xi_{1/2}(x_2) = t_1$. The four types of confidence intervals are studied for all the parameters, except that bootstrap-t intervals are not attempted for $\xi_{1/2}(\cdot)$. The bootstrap intervals are formed from bootstrap samples taken by each of the three methods of forming the bootstrap sample. Therefore ten methods are studied for β_0 and $S(\cdot)$, and seven methods are studied for $\xi_{1/2}(\cdot)$. The number of simulations in all the studies is 2000.

3.3 Summary of Results of the Monte Carlo Studies

The performance of 90% confidence intervals for each of the five parameters listed above is studied in terms of coverage probability and average length. For the confidence intervals for $\xi_{1/2}(x)$, we report median rather than mean length; this is explained when those results are discussed. The results of the studies are summarized by plots of coverage probability versus sample size and plots of average (or median) length versus sample size. In studying the results, we found that plotting the mass of numbers enabled us to make much more sense of them than simply scanning them in a table. The plots are grouped according to the type of parameter being estimated (β_0, $S(\cdot)$, or $\xi_{1/2}(\cdot)$), and are arranged in nine figures. The plots and an explanation of them are in the Appendix.

Consider β_0 first, with results plotted in Figures 1 and 2. For this parameter, we note a strong interaction effect between the two factors which determine the confidence intervals, the method of drawing the sample and the method of forming the interval from a given sample. For example, the variability among the coverage probabilities of the three methods of drawing the samples is much less for the percentile intervals than for the other two types of intervals. Also, Method 1 of drawing the sample produces the best bootstrap-t intervals, but the worst percentile intervals. Now we give the practical implications of the results for β_0. Overall, the asymptotic method seems to do the best in this study. It has close to or slightly above the nominal coverage probability throughout, yet its average length is at worst only very slightly greater than the average length of any of the nine bootstrap methods. It has slightly larger average length than the bootstrap-t intervals (henceforth referred to as boot-t intervals) but the boot-t intervals all have low coverage probability. Among the bootstrap methods, a close competitor to the asymptotic method is the hybrid interval from Method 2 samples (henceforth referred to as hybrid/Method 2). These intervals give substantially greater coverage probability than the asymptotic intervals (and greater than the nominal level), at the cost of greater average length than the asymptotic intervals. The difference in average length is substantial only for the sample sizes up to $n = 50$. We mention one other important result about the bootstrap methods here. This is that the percentile/Method 1 intervals are bad: They have larger average length than the other methods for sample sizes up to $n = 60$, yet no better, or *worse* coverage probability. The hybrid/Method 1 intervals, which are the same length as the percentile/Method 1 intervals, have very high coverage probability. So for estimation of β_0, when taking the bootstrap sample by the simplest method, the most obvious methods of forming bootstrap intervals are bad because they are too long. In the case of the percentile intervals, which are the most often used, they are not only too long, but they get the "wrong" part of the bootstrap distribution and thus have low coverage probability as well.

Figures 3–6 summarize the results for estimation of $S(\cdot)$, for two values of t. We can make a general statement here regarding the methods of forming the intervals: Roughly one could say that the coverage probabilities of the boot-t intervals are too high, the coverage probabilities of the hybrid intervals are too low, and the coverage probabilities of the percentile intervals are closest to nominal. Next we give the practical implications for this parameter. The asymptotic intervals have unacceptably low coverage probability throughout; they lose. (This fact was noted by Link (1984), who tried several transformations to improve upon the standard asymptotic method.) The winner is the percentile/Method 1 interval, with the percentile/Method 2 interval a close second.

In the case of estimation of $\xi(\cdot)$, all of the methods of forming confidence intervals produced occasional very long intervals; the distribution of the length of confidence intervals was strongly positively skewed. We noted that for the three methods of taking bootstrap samples, the most erratic fluctuation of length and largest mean length was with Method 3, then Method 2. Method 1 was the most stable. The median length provides a more relevant comparison of the three methods.

Figures 7–9 show the results for estimation of $\xi(\cdot)$, for two values of x. Rather than average length of the intervals, the median length is plotted, in Figure 9. The most striking thing about the plots is that the hybrid intervals have extremely low coverage probability. The asymptotic intervals have somewhat low coverage probability for $\xi(x_1)$, for the sample sizes of $n = 60$ and up, and they have high coverage probability for $\xi(x_2)$. Regarding the performance of the three methods of drawing bootstrap samples in conjunction with percentile intervals, Method 2 is best for estimation of $\xi(x_1)$; it has close to the nominal coverage probability and the shortest median length of the three methods. Method 3 is actually the worst of the three, at least for sample sizes up to about $n = 80$. In the case of estimation of $\xi(x_2)$, Method 3 is not such a clear loser; it has low coverage probability but also has markedly shorter median length than the other methods. Methods 1 and 2 are very close competitors here.

3.4 The Study With An Influential Point

This study consisted of only one situation, as contrasted with nineteen situations included in the main study. Therefore it is reported here simply to give an indication of the differences that arise when an influential point is included.

The data sets in this simulation study were generated as described in Section 3.1, conditional on a covariate/censoring pattern. The data set which was conditioned on was handpicked to include an influential point, and the lifetime and censoring distributions were constructed so that this handpicked data set was a possible occurrence in the model. This was done through the use of mixture distributions for the lifetime Y and the censoring time C.

To make this clear, it is perhaps best to spell out the details for the particular situation considered. The sample size was taken to be $n = 30$, and the regression parameter β_0 was taken equal to 2. Then, X_1, the first covariate value, was chosen to be unusually large. That is, twenty-nine out of the thirty covariate values were generated initially from the Uniform(0,1) distribution, whereas X_1 was taken equal to 1.5. Since the regression parameter is positive, lifetimes associated with $X = 1.5$ would tend to be much smaller than the other lifetimes. An influential point would be one with *large* X but *large* lifetime.

The lifetime distribution F was taken to be a mixture of Exponential distributions, one with mean 50 occurring with probability .04, the other with mean 1 occurring with probability .96. For the unusual lifetime distribution with mean 50, if $X = 1.5$ the mean lifelength is 2.49.

The influence of such a point is likely to be lost by censoring, so a mixture distribution on C was used to allow the possibility that such a point would remain uncensored. The censoring variable was taken equal to 3 with probability .04 and otherwise Exponential with mean 1.0, so that average percent censoring was about 25%.

In getting the initial data set as in Step 1 of Section 3.1, the lifetime associated with the large covariate value $X_1 = 1.5$ was set equal to the median of the unusual lifetime distribution (mean = 50) for that value of X, and it was forced to be uncensored.

The performance of 90% confidence intervals is summarized in Table 1 in the Appendix. One observation is that the asymptotic method is no longer best for estimation of β_0; both Method 2 and Method 3 percentile intervals beat it uniformly, that is, with higher coverage probability yet shorter average length. We also see that Method 3 of drawing the bootstrap samples does much better than it did in the other studies. We note that no one method is uniformly better than the others for all the parameters, and Methods 2 and 3 of drawing the samples are close throughout.

4 Discussion

Our first comment is that although bootstrap methods do not improve upon the asymptotic method in the case of β_0, they do compete closely with the asymptotic method for β_0 and offer substantial improvement for $S(\cdot)$ and $\xi_{1/2}(\cdot)$. For the practitioner who wants to use the bootstrap in the Cox model, results of the Monte Carlo studies suggest which bootstrap methods would be preferred among those now readily available. We would like to be able to recommend a single method appropriate for all parameters, but as mentioned earlier, this is not possible. The performance of the methods was markedly different from parameter to parameter. A summary of conclusions drawn from the Monte Carlo studies is as follows. For β_0, we recommend the hybrid/Method 2 intervals. For $S(\cdot)$ and $\xi_{1/2}(\cdot)$, we recommend the percentile/Method 2 intervals. (Method 1 did slightly better than Method 2 for $S(\cdot)$, but its advantage was so slight that it hardly seems worth the extra effort of using a different method of resampling.)

Regarding the methods of drawing the bootstrap samples, overall this factor appears to be less important than the method of forming the intervals in its effect on performance of confidence intervals. However, some effects of this factor were noted. First, the most sophisticated way of drawing the samples, Method 3, gives much more erratic results than the two simpler methods, and we would prefer to avoid this method. Also, Method 1 was unreliable for estimation of β_0, and therefore we would prefer to avoid this simplest method of resampling as well.

An important lesson learned from this work is that the choice of method of forming the interval is not one to be made lightly. There are big differences among the methods we studied in terms of coverage probabilities. In particular, the bootstrap-t was very unstable throughout, and the hybrid method had very low coverage probability for $\xi_{1/2}(x)$. Our results contrast sharply with those of Owen (1988), who included five bootstrap methods in a large-scale Monte Carlo study of confidence intervals for the mean from a random sample, and found the bootstrap-t to be a clear winner throughout.

However, it is important to realize that the present work does not provide the final word on bootstrap confidence intervals in the Cox model. Use of the iterated bootstrap refinement seems especially appropriate for the bootstrap-t intervals for β_0 and $S(\cdot)$. These intervals were erratic in length and coverage probability in that they were either very short with low coverage probability or very long with high coverage probability. However, they never lost out to another method by having both lower coverage probability and greater length; and so, this leads us to believe that the bootstrap-t interval, corrected by the iterated bootstrap method, could prove to be useful. The same comment applies to the percentile/Method 3 intervals, except for estimation of $\xi_{1/2}(x_1)$, indicating that Method 3 of sampling might prove useful if further refined, particularly in light of its good performance in the study with the influential point. Monte Carlo studies of the performance of the iterated method will only be feasible if one can develop analytic techniques to replace the inner layer of resampling in this method, as is done in DiCiccio, Martin, and Young (1990) for a simpler problem.

5 Asymptotic Validity of Method 3 of Bootstrapping

Asymptotics for the bootstrap in the Cox model have been dealt with to date in two unpublished technical reports. Gu (1991) gives arguments for the asymptotic validity of Method 1 and Hjort (1985) for asymptotic validity of Method 2. These authors deal with use of the bootstrap to approximate the distributions of the estimates of the regression parameter and of the baseline survival function (Theorem 4.2 Part 5 and Theorem 5.2 Part 2 of Gu (1991); the proposition on p. 13 of Hjort (1985)).

In the present paper, we consider Method 3 of bootstrapping. Instead of looking at the Cox model in full generality, we consider the mathematically simpler situation in which β_0 is known to be zero. This corresponds to the familiar framework for which the Kaplan-Meier estimator (henceforth KME) is the well-known and well-studied estimator of the survival function. We prove that when bootstrap samples are taken by Method 3, the standardized KME computed from the bootstrap sample converges weakly to the same Gaussian process to which the standardized KME itself converges. (In his unpublished Ph.D. dissertation, Kim (1990) gives an alternative proof of this result.)

Although the model that we consider is a special case of the Cox model, the arguments that we use can be used to prove the analogous results for the Cox model. The reason we consider the simpler case here is that it is possible to give a complete proof in a reasonable amount of space. At the end of this section, we very briefly indicate how the arguments need to be modified in order to deal with the general model.

The notation here is chosen to be as similar as possible to that used in the rest of the paper. Since there is no covariate vector X_i, associated with individual i are just a lifelength Y_i and a censoring time C_i; the observed data are $T_i = \min(Y_i, C_i)$ and $\delta_i = I(Y_i \leq C_i)$. Of course the KME's $\hat{F}$ and $\hat{G}$ of the lifetime and censoring distributions depend on the sample size n, as do the relevant stochastic processes. However, here we suppress the subscript n to simplify notation and to be consistent with the rest of the paper.

Theorem 1 *Assume the random censorship model where the failure time Y has continuous distribution function F and the censoring time C has continuous distribution function G. Let H be the distribution function of the observed time, i.e. $H = 1-(1-F)(1-G)$. Let $\hat{F}^*$ be the Kaplan-Meier estimate of $\hat{F}$ computed from the data resampled by Bootstrap Method 3. Then as $n \to \infty$*

$$\sqrt{n}\left(\frac{\hat{F}^* - \hat{F}}{1-\hat{F}}\right) \xrightarrow{d} W \quad in\ D[0,\tau] \quad \text{a.s.} \tag{5.1}$$

for any $\tau < \sup\{t : H(t) < 1\}$, where W is a mean zero Gaussian process with independent increments and variance function given by

$$Var(W(t)) = \int_{[0,t]} \frac{1}{(1-F(s))(1-H_-(s))} dF(s). \tag{5.2}$$

The notation "a.s." in (5.1) signifies that the weak convergence result holds along almost every infinite sequence $(T_1, \delta_1), (T_2, \delta_2), \ldots$.

Remark The assumption that F and G be continuous is actually superfluous (we discuss this at the end of the proof); we use the notation F_- and G_- so that the formulas will continue to be valid if F and G are not continuous.

Proof We rely heavily on Gill (1980), who establishes weak convergence of the KME for the nonbootstrapped case, using the machinery of counting processes and martingale central limit theorems. A good reference for this material is the first two chapters of Fleming and Harrington (1990). To prove Theorem 1, we apply the arguments of Gill (1980) conditional on the particular infinite sequence $\{(T_i, \delta_i);\ i = 1, 2, \ldots\}$ (we shall see below that the weak convergence statement in (5.1) actually holds along every sequence such that $\hat{F} \to F$ and $\hat{G} \to G$).

For Bootstrap Method 3, we first resample $Y_i^* \sim \hat{F}$, $C_i^* \sim L_i$, where L_i is represented as

$$L_i(t) = I(\delta_i = 0)I(t \in [T_i, \infty)) + I(\delta_i = 1)\frac{(\hat{G}(t) - \hat{G}(T_i))I(t \in [T_i, \infty))}{1 - \hat{G}(T_i)}. \tag{5.3}$$

We then form $T_i^* = \min(Y_i^*, C_i^*)$ and $\delta_i^* = I(Y_i^* \le C_i^*)$, for $i = 1, \ldots, n$. (For the sake of definiteness, $\hat{G}$ in (5.3) is taken to be 1 at the largest observation and beyond, whether the δ corresponding to this observation is 0 or 1.)

Our situation is complicated by the following three factors.

1 The survival distribution function $\hat{F}$ varies with n.

2 The function $\hat{F}$ is discontinuous. Thus the standard counting processes associated with the pairs (T_i, δ_i) may jump at the same time, so that they do not form a "multivariate counting process".

3 The censoring distribution function L_i varies with i as well as with n.

We refer to the standardized version of $\hat{F}^*$ as Z^*; that is,

$$Z^*(t) = \sqrt{n}\Big(\frac{\hat{F}^*(t) - \hat{F}(t)}{1 - \hat{F}(t)}\Big). \tag{5.4}$$

Let

$$F^\dagger(t) = \hat{F}(t \wedge T^*_{(n)}), \tag{5.5}$$

where $T^*_{(n)} = \max\{T^*_1, \ldots, T^*_n\}$, and let

$$Q^*(t) = \sqrt{n}\Big(\frac{\hat{F}^*(t) - F^\dagger(t)}{1 - F^\dagger(t)}\Big). \tag{5.6}$$

The proof of Theorem 1 proceeds in roughly the following three steps.

1 We show that the process $\{Q^*(t),\ t \in [0,\tau]\}$ can be represented as the stochastic integral of a predictable process with respect to a martingale and is therefore a martingale.

2 We show that the martingale Q^* satisfies Conditions (I) of Theorem 4.2.1 of Gill (1980), which gives weak convergence of Q^* to the process W with variance function given by (5.2).

3 We show that the processes Z^* and Q^* are asymptotically equivalent.

Let $\hat{\Lambda}(t)$ be the cumulative hazard function associated with $\hat{F}$, i.e.

$$\hat{\Lambda}(t) = \int_{[0,t]} \frac{d\hat{F}(s)}{1 - \hat{F}_-(s)}.$$

For $i = 1, \ldots, n$, define the processes

$$\begin{aligned}
N_i^*(t) &= I(T_i^* \le t,\ \delta_i^* = 1), \\
N_i^{*U}(t) &= I(T_i^* \le t,\ \delta_i^* = 0), \\
V_i^*(t) &= I(T_i^* \ge t), \\
A_i^*(t) &= \int_{[0,t]} V_i^*(s) d\hat{\Lambda}(s), \\
M_i^*(t) &= N_i^*(t) - A_i^*(t).
\end{aligned}$$

Define further, for each i, the filtration

$$\mathcal{F}_t^{(i)} = \sigma((N_i^*(s), N_i^{*U}(s));\ s \le t), \quad t \in [0,\tau].$$

It is well known that with respect to the filtration $\mathcal{F}_t^{(i)}$, M_i^* is a martingale with predictable variation process

$$< M_i^*, M_i^* > (t) = A_i^*(t) \tag{5.7}$$

(cf. Theorem 1.3.1 and Theorem 2.5.2 Part 1 of Fleming and Harrington (1990)). Let $\mathcal{F}_t$ be the join of $\mathcal{F}_t^{(i)}$ for $i = 1, \ldots, n$, i.e.

$$\mathcal{F}_t = \sigma((N_i^*(s), N_i^{*U}(s));\ s \le t;\ i = 1, \ldots, n).$$

Then since the pairs (Y_i, C_i) are independent, it is easy to see that with respect to the filtration $\mathcal{F}_t$, the M_i^*'s are still martingales and their predictable variation processes are still given by (5.7) (see for example Lemma A.1 of Doss and Chiang (1990)); moreover, with respect to $\mathcal{F}_t$, the M_i^*'s are orthogonal (Lemma 2.6.1. of Fleming and Harrington (1990)).

We now proceed to show that Q^* is a martingale. Define the processes

$$\begin{aligned} M^*(t) &= \sum_{i=1}^{n} M_i^*(t), \\ V^*(t) &= \sum_{i=1}^{n} V_i^*(t), \\ J^*(t) &= I(V^*(t) > 0), \end{aligned} \tag{5.8}$$

Let τ_n be any number such that $\hat{F}(\tau_n) < 1$. We then have the identity

$$Q^*(t) = \sqrt{n} \int_{[0,t]} \frac{1 - \hat{F}_-^*(s)}{1 - \hat{F}(s)} \frac{J^*(s)}{V^*(s)} \, dM^*(s) \quad \text{for } t \in [0, \tau_n], \tag{5.9}$$

which follows from Equation (3.2.13) of Gill (1980). (Here and throughout, 0/0 is defined to be 0.) Because the integrand in (5.9) is $\mathcal{F}_t$-predictable, we see from (5.9) and the fact that M^* is a martingale, that Q^* is the stochastic integral of a predictable process with respect to a martingale, and is therefore a martingale.

We are now in a position to apply Theorem 4.2.1 of Gill (1980), which gives conditions that entail weak convergence of processes that are stochastic integrals with respect to M^*. This theorem is general enough to accommodate our situation. It requires implicitly that Assumption 3.1.1 of Gill (1980) hold, i.e. that there exist a filtration with respect to which the processes M_i^*'s are orthogonal martingales, with predictable variation processes given by (5.7). We have just shown that this structure holds. To use the theorem, we need to verify Conditions (I) of Theorem 4.2.1 of Gill (1980). For the infinite sequence $(T_1, \delta_1), (T_2, \delta_2), \ldots$, these conditions are as follows.

A $\hat{F}$ converges uniformly on $[0, \tau]$ to F as $n \to \infty$; $\Lambda = \int \frac{1}{1-F_-} \, dF$ is finite on $[0, \tau]$.

B There is a function h that is left continuous with right-hand limits and is of bounded variation on $[0, \tau]$ such that $\left(\frac{1-\hat{F}_-^*}{1-\hat{F}}\right)^2 \frac{n}{V^*} J^*$ converges uniformly on $[0, \tau]$ in bootstrap probability to h.

C $V^*(t) \to \infty$ in bootstrap probability as $n \to \infty$ for each $t \in [0, \tau]$.

We shall show that Conditions A, B, and C are satisfied for any infinite sequence $(T_1, \delta_1), (T_2, \delta_2), \ldots$ for which

$$\sup_{0 \le t \le \tau} |\hat{F}(t) - F(t)| \to 0 \quad \text{and} \quad \sup_{0 \le t \le \tau} |\hat{G}(t) - G(t)| \to 0. \tag{5.10}$$

We note that the set of such sequences has probability one by the results of Földes, Rejtö and Winter (1980).

So from now on we assume that the infinite sequence $(T_1, \delta_1), (T_2, \delta_2), \ldots$ satisfies (5.10). Let E_*, Var_*, P_*, and $\xrightarrow{\mathcal{P}_*}$ denote expectation, variance, probability, and convergence in probability under bootstrapping, respectively, i.e. these are taken conditional on the sequence $(T_1, \delta_1), (T_2, \delta_2), \ldots$. The first part of Condition A is then obviously satisfied and the second part follows from the definition of τ. We will show that Condition B is satisfied for h given by

$$h = \left(\frac{1 - F_-}{1 - F}\right)^2 \frac{1}{1 - H_-}. \tag{5.11}$$

If F and G are continuous, then h is left continuous with right limits and is of bounded variation on $[0, \tau]$.

We deal first with the term n/V^* in Condition B, where $V^*(t) = \sum_{i=1}^n I(T_i^* \geq t)$. Note that $E_*(V^*(t)) = \sum_{i=1}^n (1 - \hat{F}_-)(1 - L_{i-})$. We will need the following lemma, which gives a surprising connection between the L_i's and $\hat{G}$.

Lemma 1 *For L_i given by (5.3) we have*

$$\frac{1}{n} \sum_{i=1}^n L_i(t) = \hat{G}(t).$$

Proof of Lemma 1 Let $T'_{(1)} < T'_{(2)} < \cdots < T'_{(m)}$ be the distinct ordered observations. Without loss of generality we can assume that censored and uncensored observations are not tied. If they are, censored times are considered to have occurred just after uncensored times, following the usual convention. Let $\delta'_{(1)} < \delta'_{(2)} < \cdots < \delta'_{(m)}$ be the corresponding indicator variables and let $v'_{(1)} < v'_{(2)} < \cdots < v'_{(m)}$ be the number of ties. Writing $1 = I(\delta_i = 0,\ T_i \leq t) + I(\delta_i = 1,\ T_i \leq t) + I(T_i > t)$, we have

$$\begin{aligned}
1 - \frac{1}{n} \sum_{i=1}^n L_i(t) &= \frac{1}{n} \sum_{i=1}^n \Big(I(\delta_i = 0,\ T_i \leq t) + I(\delta_i = 1,\ T_i \leq t) + I(T_i > t) \Big) \\
&\quad - \frac{1}{n} \sum_{i=1}^n I(\delta_i = 0,\ T_i \leq t) \\
&\quad - \frac{1}{n} \sum_{i=1}^n \frac{\hat{G}(t) - \hat{G}(T_i)}{\hat{G}(T_i)} I(\delta_i = 1,\ T_i \leq t) \\
&= \frac{1}{n} \Big(\sum_{i=1}^n I(T_i > t) + \sum_{i=1}^n \frac{\hat{G}(t)}{\hat{G}(T_i)} I(\delta_i = 1,\ T_i \leq t) \Big) \\
&= \frac{1}{n} \Big(\sum_{i=1}^n I(T_i > t) + \sum_{i:T'_i \leq t} \frac{\hat{G}(t)}{\hat{G}(T'_i)} \delta'_{(i)} v'_{(i)} \Big).
\end{aligned}$$

Since $\hat{G}(t)$ is a self-consistent estimator we have

$$\hat{G}(t) = \frac{1}{n} \Big(\sum_{i=1}^n I(T_i > t) + \sum_{i:T'_i \leq t} \frac{\hat{G}(t)}{\hat{G}(T'_i)} \delta'_{(i)} v'_{(i)} \Big).$$

See for example Miller (1981, pp. 52–57) for the definition of the self-consistency property and a proof that it holds for the KME. His proof is for the case of no ties but can be easily modified for the case of ties with the above notation. Hence we have

$$1 - \frac{1}{n} \sum_{i=1}^n L_i(t) = \hat{G}(t),$$

and this proves the lemma.

From the lemma we see that $E_*(V^*(t)/n) = 1 - \hat{H}_-(t)$ and $\mathrm{Var}_*(V^*(t)/n) \to 0$. Therefore,

$$\left| \frac{V^*(t)}{n} - (1 - \hat{H}_-(t)) \right| \xrightarrow{\mathcal{P}_*} 0 \tag{5.12}$$

for each t. As explained on p. 70 of Gill (1980), the results of van Zuijlen (1978) give that the convergence in (5.12) is uniform over $\mathbf{R}$. By (5.10), $\hat{H}(t) \to H(t)$ uniformly on $[0, \tau]$, and this implies that

$$\left| \frac{n}{V^*(t)} - \frac{1}{1 - H_-(t)} \right| \xrightarrow{\mathcal{P}_*} 0 \quad \text{uniformly for } t \in [0, \tau]. \tag{5.13}$$

By Theorem 4.1.1 of Gill (1980) we have $\sup_{0 \leq t \leq \tau} |\hat{F}^*(t) - \hat{F}(t)| \xrightarrow{\mathcal{P}_*} 0$. This, together with (5.10) gives

$$\frac{1 - \hat{F}^*_-}{1 - \hat{F}} \xrightarrow{\mathcal{P}^*} \frac{1 - F_-}{1 - F} \quad \text{uniformly for } t \in [0, \tau]. \tag{5.14}$$

Let us now consider $J^*(t)$ defined by (5.8). We have

$$\begin{aligned}
P_*(J^*(t) \neq 1 \text{ for some } t \in [0, \tau]) &= \prod_{i=1}^{n} P_*(T_i^* < \tau) \\
&= \prod_{i=1}^{n} \Big(1 - P_*(T_i^* \geq \tau)\Big) \\
&\leq \prod_{i=1}^{n} \exp\Big(-\hat{\bar{F}}_-(\tau)(1 - L_{i-}(\tau))\Big) \qquad (5.15) \\
&= \exp\Big(-n\hat{\bar{F}}_-(\tau)\hat{\bar{G}}_-(\tau)\Big) \\
&= \exp\Big(-n\hat{\bar{H}}_-(\tau)\Big) \\
&\to 0,
\end{aligned}$$

where in the above the third line follows from the inequality $1 - x \leq \exp(-x)$ and the fourth follows from Lemma 1. This, combined with (5.13) and (5.14), shows that Condition B is satisfied for any sequence $(T_1, \delta_1), (T_2, \delta_2), \ldots$ satisfying (5.10). Condition C is trivially satisfied by (5.12).

We can now apply Theorem 4.2.1 of Gill (1980) to conclude that (5.1) holds if $\sqrt{n}\left(\frac{\hat{F}^* - \hat{F}}{1 - \hat{F}}\right)$ is replaced by Q^*. (Note that since as $n \to \infty$ we have $\hat{F}(\tau) \to F(\tau) < 1$, the representation (5.9) is valid over $[0, \tau]$ for large n.) To see that the difference between Z^* and Q^* is negligible, we note that

$$P_*\Big\{ \sup_{0 \leq t \leq \tau} |Q^*(t) - Z^*(t)| \neq 0 \Big\} = P_*\{T^*_{(n)} < \tau\} \to 0$$

by (5.15). This concludes the proof of Theorem 1.

For the case in which F and G have discontinuities, the same result can be proved by the arguments of Akritas (1986, p. 1037). If we look carefully at those arguments, we see that the distributions of the censoring variables C_i^* has no role, so that the arguments can be applied directly to our situation.

Let us now return to the Cox model, and suppose that Conditions A–D of AG are satisfied. Gu (1991) and Hjort (1985) show that if $\hat{\beta}^*$ and $\hat{S}^*(\cdot)$ are the estimates of the regression parameter and baseline survival function computed from the data resampled by Method 1 or 2, then with probability one, $\sqrt{n}(\hat{\beta}^* - \hat{\beta}, \hat{S}^*(\cdot) - \hat{S}(\cdot))$ converges in distribution to the same process to which the process $\sqrt{n}(\hat{\beta} - \beta, \hat{S}(\cdot) - S(\cdot))$ converges. Gu deals with Method 1 and Hjort with Method 2. Their proofs rely heavily on the machinery developed in AG. They both focus effort on establishing second-order properties, whereas here we are concerned with first-order results only. To prove the same result for Bootstrap Method 3 one proceeds in a similar way. The heart of the proof involves checking Condition B ("asymptotic stability") and this is done via an analogue of Lemma 1. Unfortunately, the details needed to give a rigorous proof are lengthy and not straightforward, and this is the reason we have limited our result to the case where the regression parameter is known to be 0.

Acknowledgement

I thank Hani Doss for helpful discussions in connection with the proof of Theorem 1.

References

Abramovitch, L. and Singh, K. (1985), "Edgeworth Corrected Pivotal Statistics and the Bootstrap," *The Annals of Statistics*, 13, 116 - 132.

Akritas, M. G. (1986), "Bootstrapping the Kaplan-Meier Estimator," *Journal of the American Statistical Association*, 81, 1032 - 1038.

Andersen, P. K. and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, 10, 1100 - 1120.

Beran, R. (1987), "Prepivoting to Reduce Level Error of Confidence Sets," *Biometrika*, 74, 457 - 468.

Breslow, N.E. (1972), "Contribution to the Discussion of the Paper by D. R. Cox," *Journal of the Royal Statistical Society*, Ser. B, 34, 216 - 217.

Breslow, N.E. (1974), "Covariance Analysis of Censored Survival Data," *Biometrics*, 30, 89 - 100.

Burr, D. and Doss, H. (1991), "Confidence Bands for the Median Survival Time as a Function of the Covariates in the Cox Model," Technical Report 443, Stanford University, Dept. of Statistics.

Cleveland, W. S. (1985), *The Elements of Graphing Data*, Pacific Grove, CA: Wadsworth and Brooks/Cole.

Cox, D.R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187 - 220.

Dabrowska, D. M. and Doksum, K. A. (1987), "Estimates and Confidence Intervals for Median and Mean Life in the Proportional Hazard Model," *Biometrika*, 74, 799 - 807.

DiCiccio, T. J., Martin, M. A., and Young, G. A. (1990), "Analytical Approximations for Iterated Bootstrap Confidence Intervals," Technical Report 361, Stanford University, Dept. of Statistics.

Doss, H. and Chiang, Y.-C. (1991), "Choosing the Resampling Scheme When Bootstrapping: A Case Study in Reliability," Technical Report 844, Florida State University, Dept. of Statistics.

Doss, H. and Gill, R. D. (1989), "A Method for Obtaining Weak Convergence Results for Quantile Processes, With Applications to Censored Survival Data," Technical Report M-806, Florida State University, Dept. of Statistics.

Efron, B. (1981), "Censored Data and the Bootstrap," *Journal of the American Statistical Association*, 76, 312 - 319.

Efron, B. (1990), "Six Questions Raised by the Bootstrap," Technical Report 139, Stanford University, Dept. of Statistics.

Efron, B. and Tibshirani, R. (1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy" (with discussion), *Statistical Science*, 1, 54 - 77.

Fleming, T. R. and Harrington, D. P. (1991), *Counting Processes and Survival Analysis*, New York: Wiley.

Freedman, D. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1218 - 1228.

Gill, R. D. (1980), *Censoring and Stochastic Integrals*, Mathematical Centre Tract 124, Amsterdam: Mathematisch Centrum.

Gill, R. D. and Johansen, S. (1990), "A Survey of Product-integration With a View Toward Application in Survival Analysis," *The Annals of Statistics*, 18, 1501 - 1555.

Gu, Minggao (1991), "On the Edgeworth Expansion and Bootstrap Approximation for the Cox Regression Model Under Random Censorship," Technical Report, McGill University, Dept. of Mathematics and Statistics.

Hall, Peter (1988), "Theoretical Comparison of Bootstrap Confidence Intervals" (with discussion), *The Annals of Statistics*, 16, 927 - 985.

Hall, P and Martin, M. A. (1988), "Exact Convergence Rate of Bootstrap Quantile Variance Estimator," *Probability Theory and Related Fields*, 80, 261 - 268.

Hjort, N. (1985), "Bootstrapping Cox's Regression Model," Technical Report 241, Stanford University, Dept. of Statistics.

Kalbfleisch, J. D. and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley.

Kim, J. (1990), "Conditional Bootstrap Methods for Censored Data," Ph.D. dissertation, Florida State University, Dept. of Statistics.

Link, C. L. (1984), "Confidence Intervals for the Survival Function Using Cox's Proportional-hazard Model With Covariates," *Biometrics*, 40, 601 - 610.

Lo, S.-H. and Singh, K. (1986), "The Product-limit Estimator and the Bootstrap: Some Asymptotic Representations," *Probability Theory and Related Fields*, 71, 455 - 465.

Martin, M. A. (1990), "On Bootstrap Iteration for Coverage Correction in Confidence Intervals," *Journal of the American Statistical Association*, 85, 1105 - 1118.

Miller, R. G. (1981), *Survival Analysis*, New York: Wiley.

Owen, A. (1988), "Small Sample Central Confidence Intervals for the Mean," Technical Report 302, Stanford University, Dept. of Statistics.

Peto, R. (1972). "Discussion on Professor Cox's Paper," *Journal of the Royal Statistical Society*, Ser B, 34, 205 - 207.

Ramlau-Hansen, H. (1983). "Smoothing Counting Process Intensities by Means of Kernel Functions," *The Annals of Statistics*. 11, 453 - 466.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Singh, K. (1981), "On the Asymptotic Accuracy of Efron's Bootstrap," *The Annals of Statistics*, 9, 1187 - 1195.

Tsiatis, A. (1981), "A Large Sample Study of Cox's Regression Model," *The Annals of Statistics*, 9, 93 - 108.

Van Zuijlen, M. C. A. (1978), "Properties of the Empirical Distribution Function for Independent Nonidentically Distributed Random Variables," *The Annals of Probability*, 6, 250 - 266.

Whittemore, A. and Keller, J. (1986), "Survival Estimation Using Splines," *Biometrics*, 42, 495 - 506.

Wu, C.F.J. (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis" (with discussion), *The Annals of Statistics*, 14, 1261 - 1350.

Appendix

This appendix contains the plots referred to in Section 3.3, and Table 1 referred to in Section 3.4. Here we explain in some detail the construction of the plots. The plots summarize the results of 19 (nineteen) simulation studies, all from the same main set-up:

1 F_X, the covariate distribution, is Uniform(0,1).

2 F, the lifetime distribution, is standard Exponential.

3 The Cox regression parameter β_0 is 2.

4 The form of the censoring distribution G is Uniform.

5 The average percent censoring is 55% (The mean of G is .25).

The 19 studies are for the sample sizes $n = 30, 40, 50, 60, 70, 80, 90, 100$, with three distinct covariate/censoring patterns for the sample sizes $n = 30, 40, 50$ and two distinct covariate/censoring patterns for the sample sizes $n = 60, \ldots, 100$.

The results are summarized with plots of coverage probability versus sample size and plots of average (or median) length versus sample size. We found in studying the results that plotting the mass of numbers enabled us to make much more sense of them than simply scanning them in a table.

Consider first the results we need to summarize for β_0 and $S(\cdot)$. There are ten types of confidence intervals studied, the asymptotic and nine types of bootstrap intervals. A bootstrap interval is determined by two factors–the method of sampling and the method of forming the interval from a given sample.

For plotting the coverage probabilities of the ten methods in the nineteen studies, we would like to be able to compare the results for the ten methods in a single plot of coverage probability versus sample size, but this produced too much overlap: We could not distinguish readily the symbols for the ten different methods. So, here we use three different plots, corresponding to the three methods of forming the intervals from a given sample. Thus the plot labelled "Percentile," for instance, contains coverage probabilities for three bootstrap methods: Percentile/Method 1, Percentile/Method 2, and Percentile/Method3. The coverage probabilities for the Asymptotic method are included on each of the three plots, so that this method can easily be compared to any bootstrap method and also as a reference to enable easier comparison of the results on the three different plots. The horizontal line at coverage probability .90 also enables easier comparison among the three plots. Two additional reference lines are drawn at coverage $.90 \pm 2(.90 \times .10/2000)^{1/2}$. If a method has exact coverage .90 then roughly 95% of the observed coverages should lie within the band formed by these lines.

With only four types of points to follow in one plot, there is relatively little problem with overlap. Here we are adhering to graphical principles put forth in Cleveland (1985, Ch. 3).

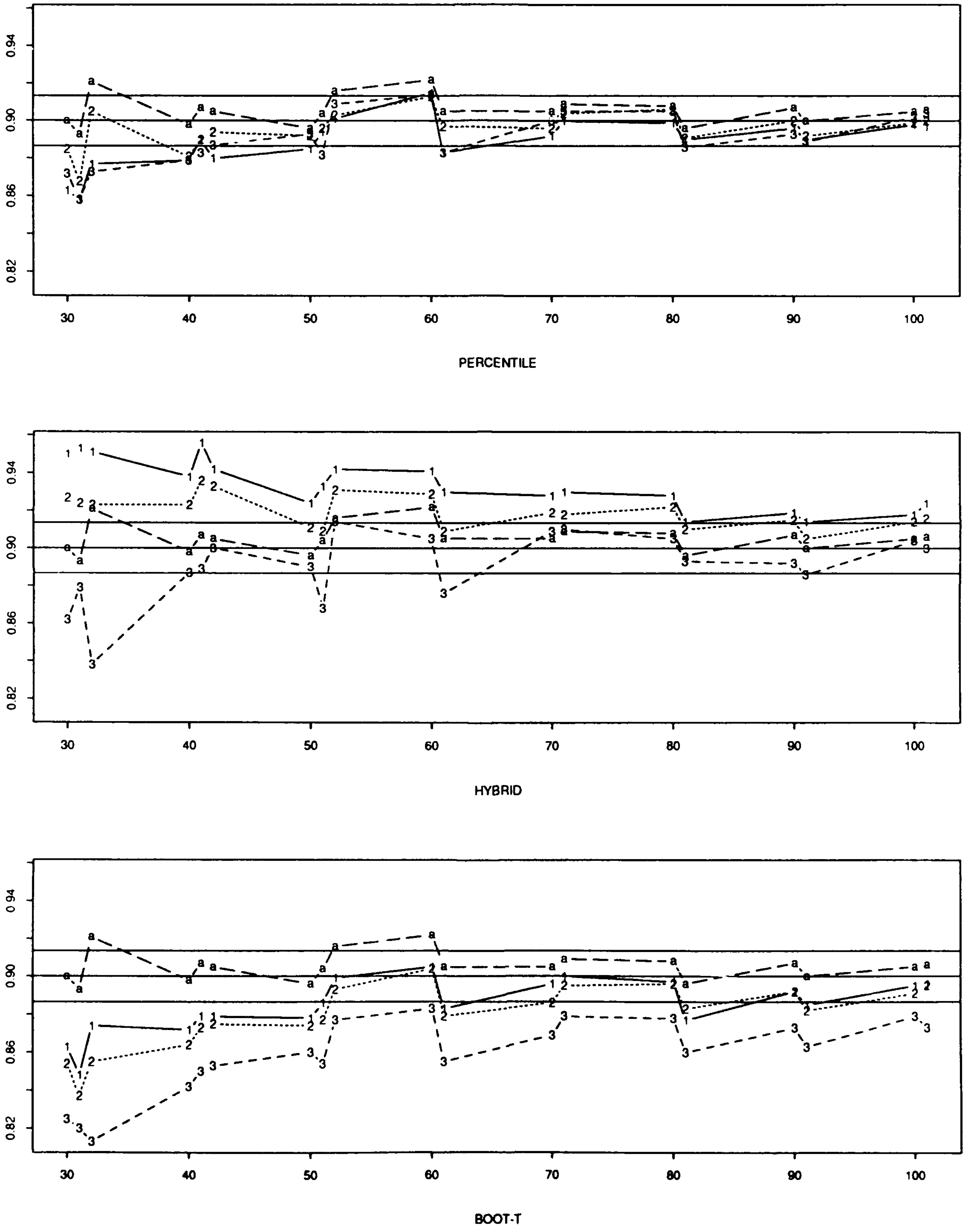
The horizontal axis in these plots is labelled sample size. However, at each sample size we had either three or two separate studies. We wanted to be able to distinguish the results of the different studies. Therefore, for the first covariate/censoring pattern at $n = 30$, we have plotted the coverage probabilities for the four methods exactly vertically aligned above the sample size $n = 30$; for the second covariate/censoring pattern at

$n = 30$, we have plotted the coverage probabilities for the four methods exactly vertically aligned, slightly to the right of $n = 30$, and so forth. We have used lines to connect symbols for the same method in different studies, both to enable the reader to more easily see the trend with sample size, and to aid the reader in identifying all points from the same study. That is, for example, without the lines it may be difficult to judge which points come from the second covariate/censoring pattern for $n = 30$. So this plot is a connected symbol graph, a kind of graph which is often used in time series but which has other uses as well; see Cleveland (1985, p. 181 and pp. 188-189).

The lengths of the percentile and hybrid intervals are the same, so there are only two plots of average length for β_0 and $S(\cdot)$. Since bootstrap-t intervals were not attempted for $\xi_{1/2}(\cdot)$, there are only two plots of coverage probability for $\xi_{1/2}(\cdot)$ and one plot of median length. For each of the five parameters studied, there are two figures, one for coverage probability and one for average length. The only exception to this is that the two plots of median length for the two values of x at which $\xi_{1/2}(x)$ is studied are put into a single figure, Figure 9.

Figure 1: Coverage Probability vs. Sample Size for Beta

1 = Method 1, 2 = Method 2, 3 = Method 3, a = Asymptotic



PERCENTILE

HYBRID

BOOT-T

Figure 2: Average Length vs. Sample Size for Beta

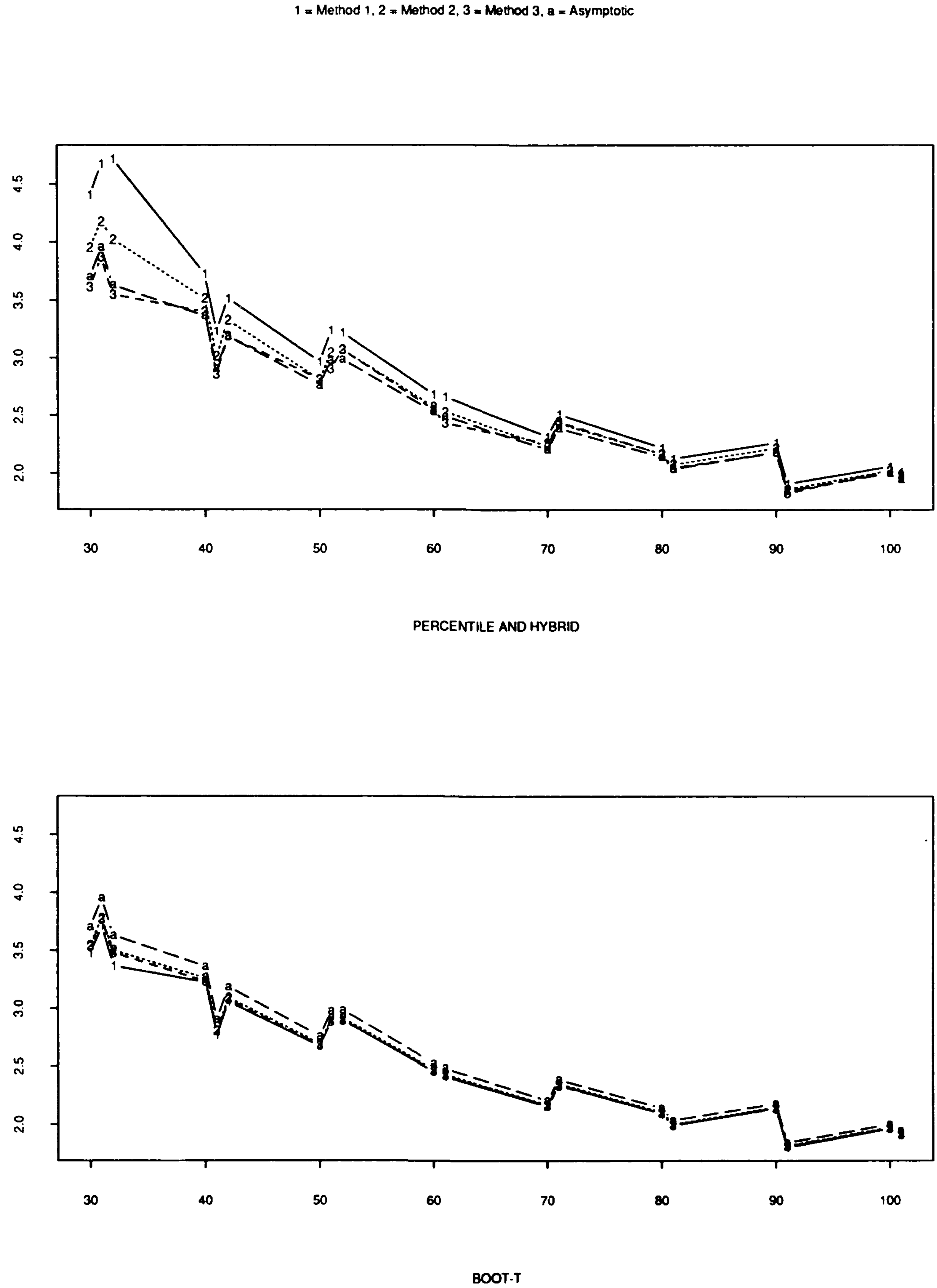


Figure 3: Coverage Probability vs. Sample Size for S(.106)

1 = Method 1, 2 = Method 2, 3 = Method 3, a = Asymptotic

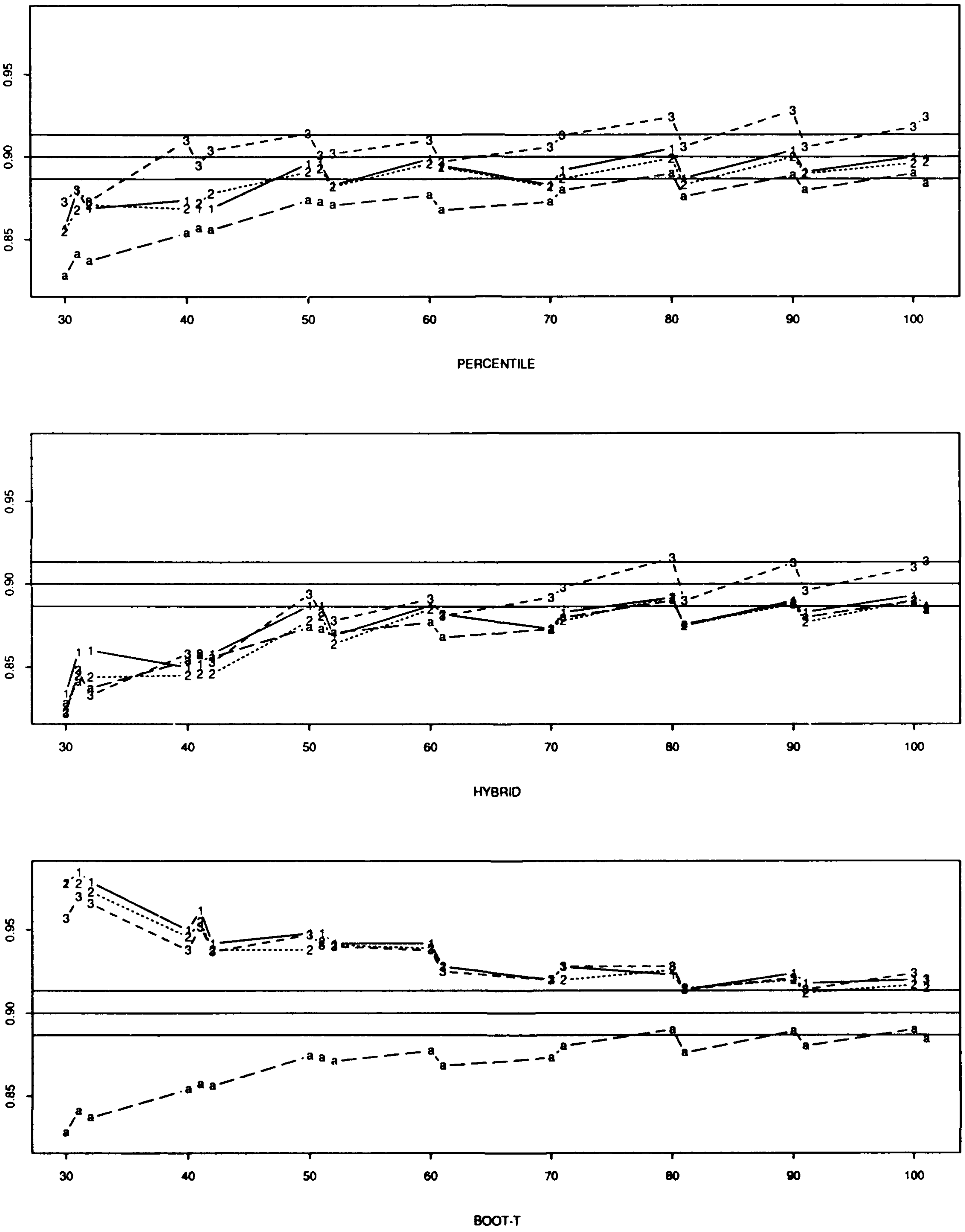
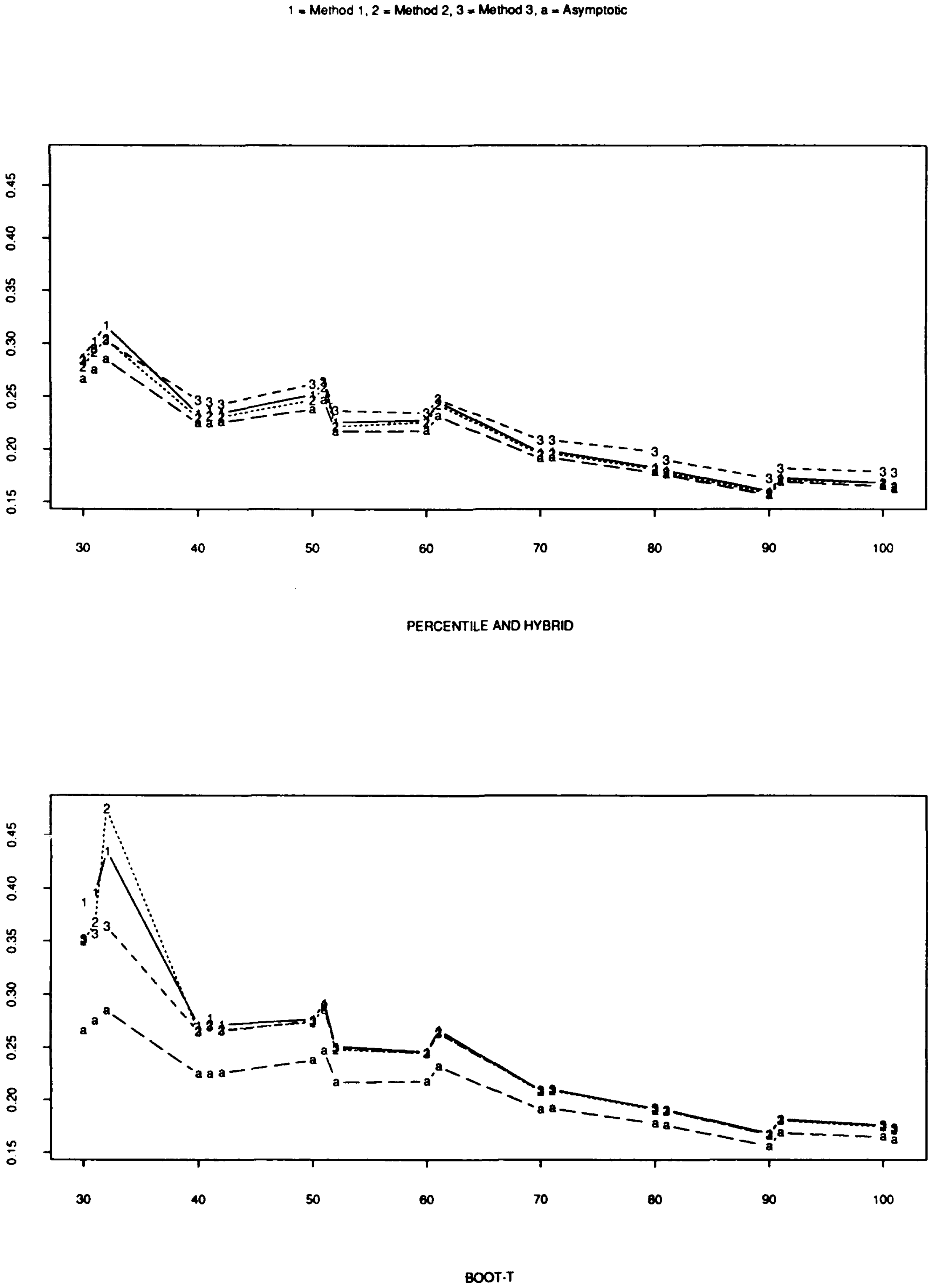


Figure 4: Average Length vs. Sample Size for S(.106)

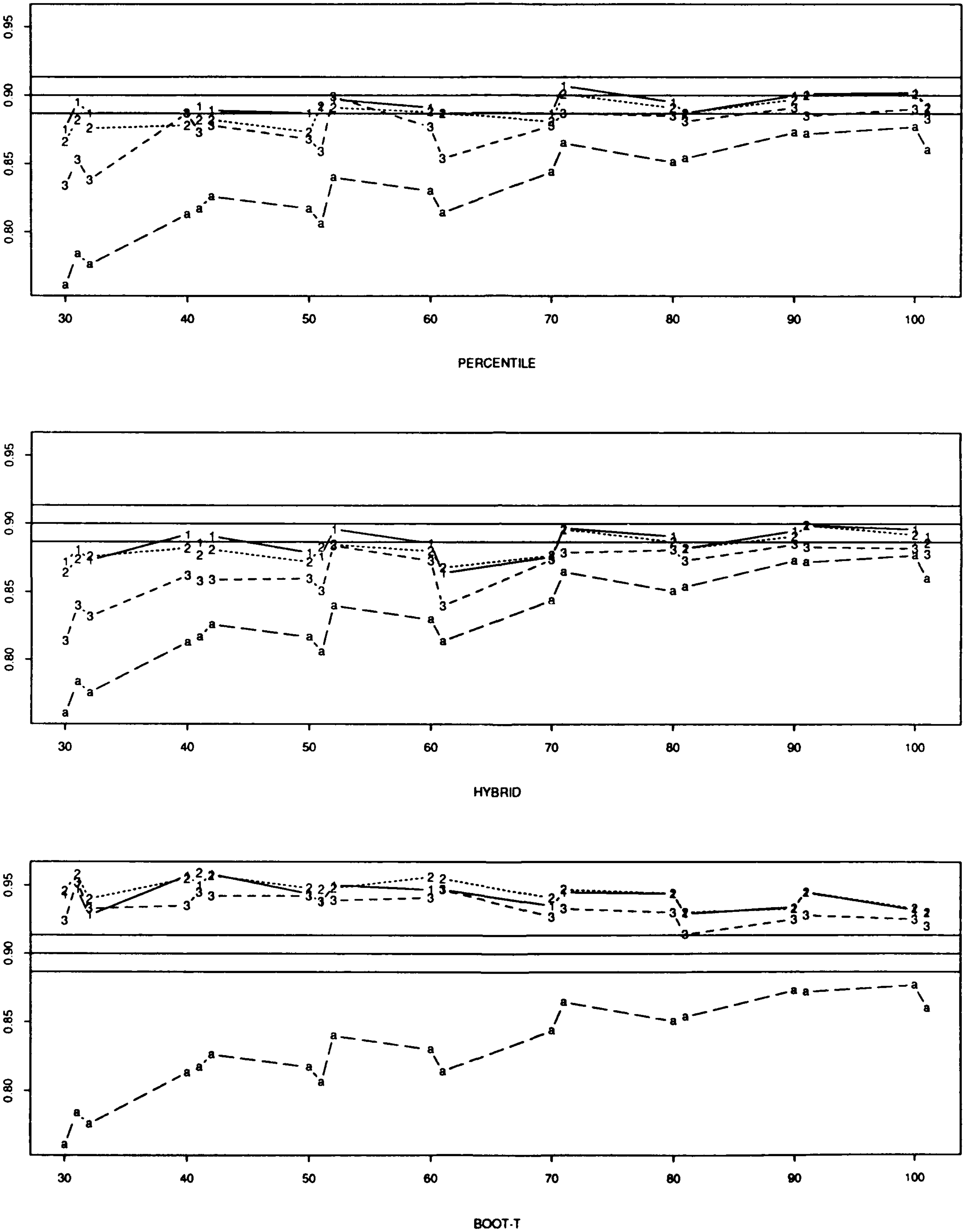


PERCENTILE AND HYBRID

BOOT-T

Figure 5: Coverage Probability vs. Sample Size for S(.255)

1 = Method 1, 2 = Method 2, 3 = Method 3, a = Asymptotic

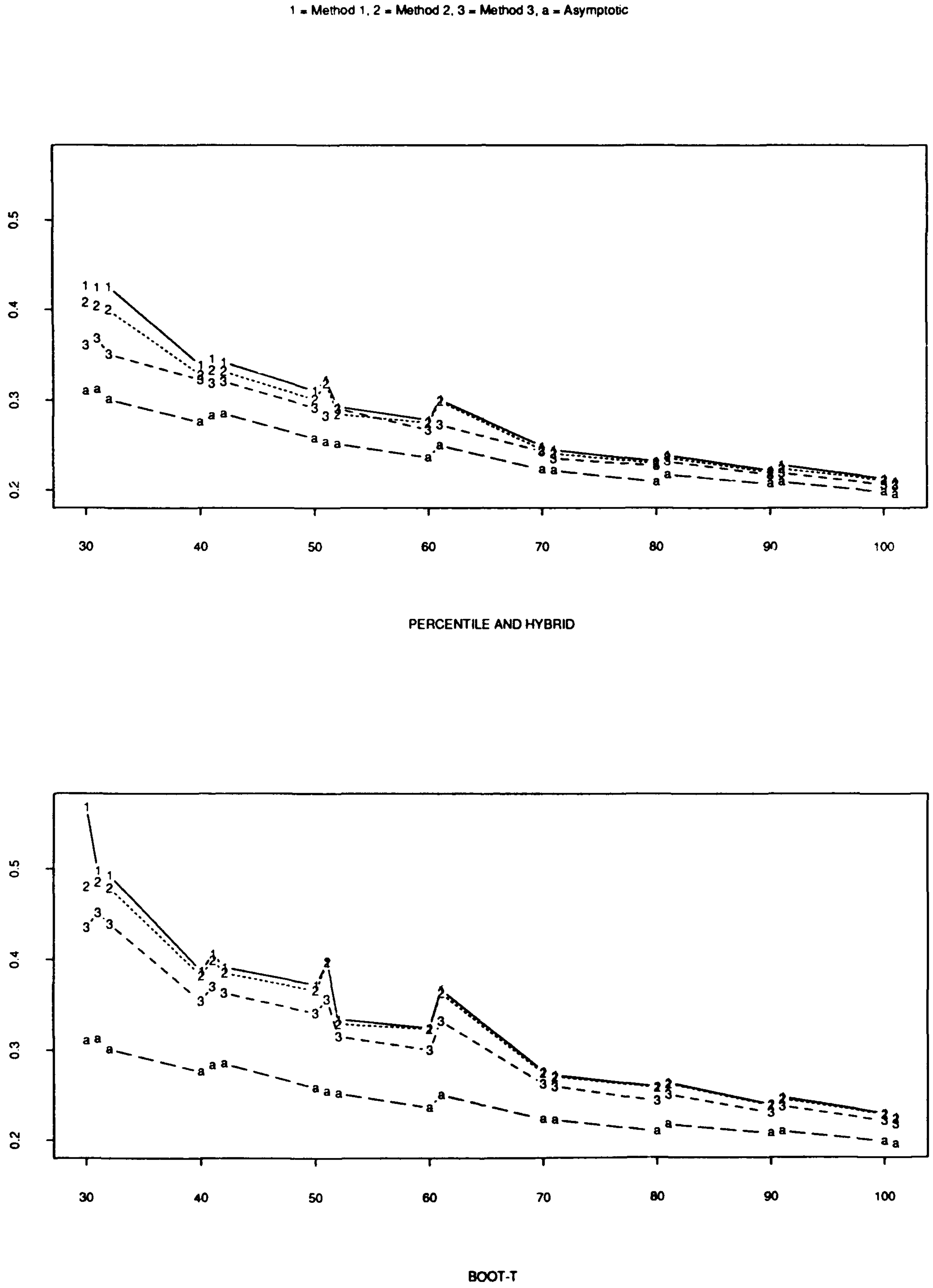


PERCENTILE

HYBRID

BOOT-T

Figure 6: Average Length vs. Sample Size for S(.255)



PERCENTILE AND HYBRID

BOOT-T

Figure 7: Coverage Probability vs. Sample Size for Median Survival at X = .5

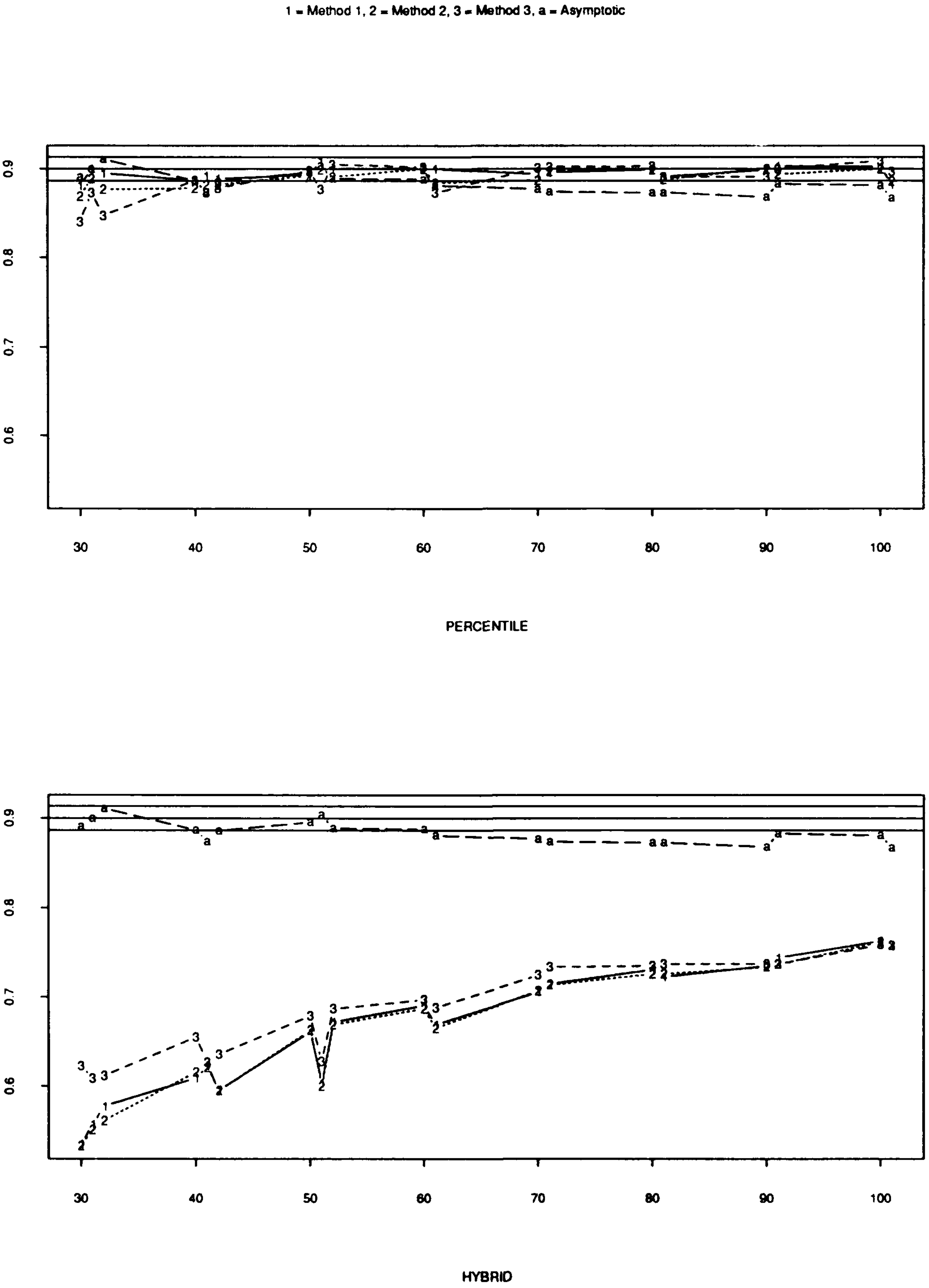


Figure 8: Coverage Probability vs. Sample Size for Median Survival at X = .939

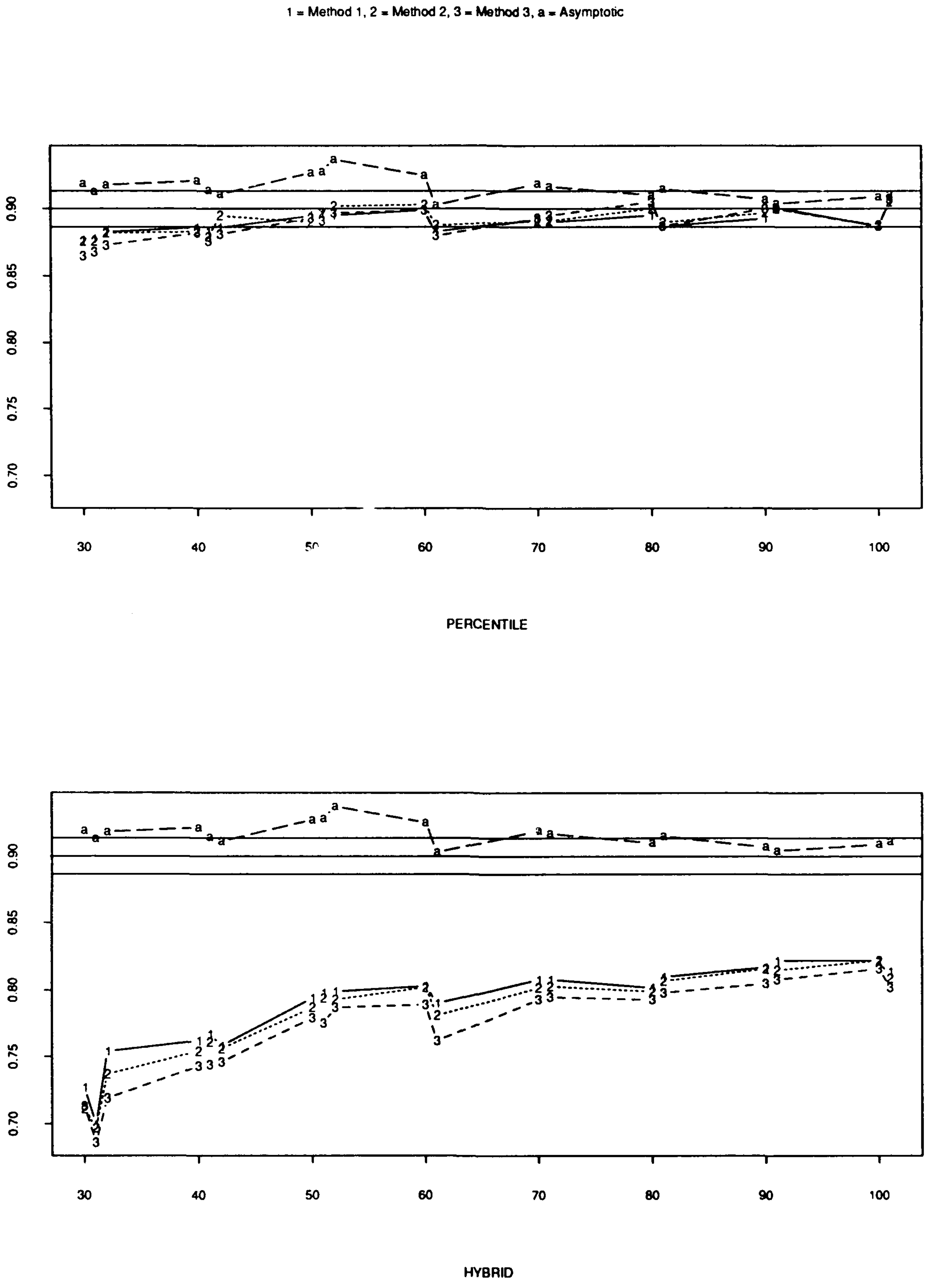
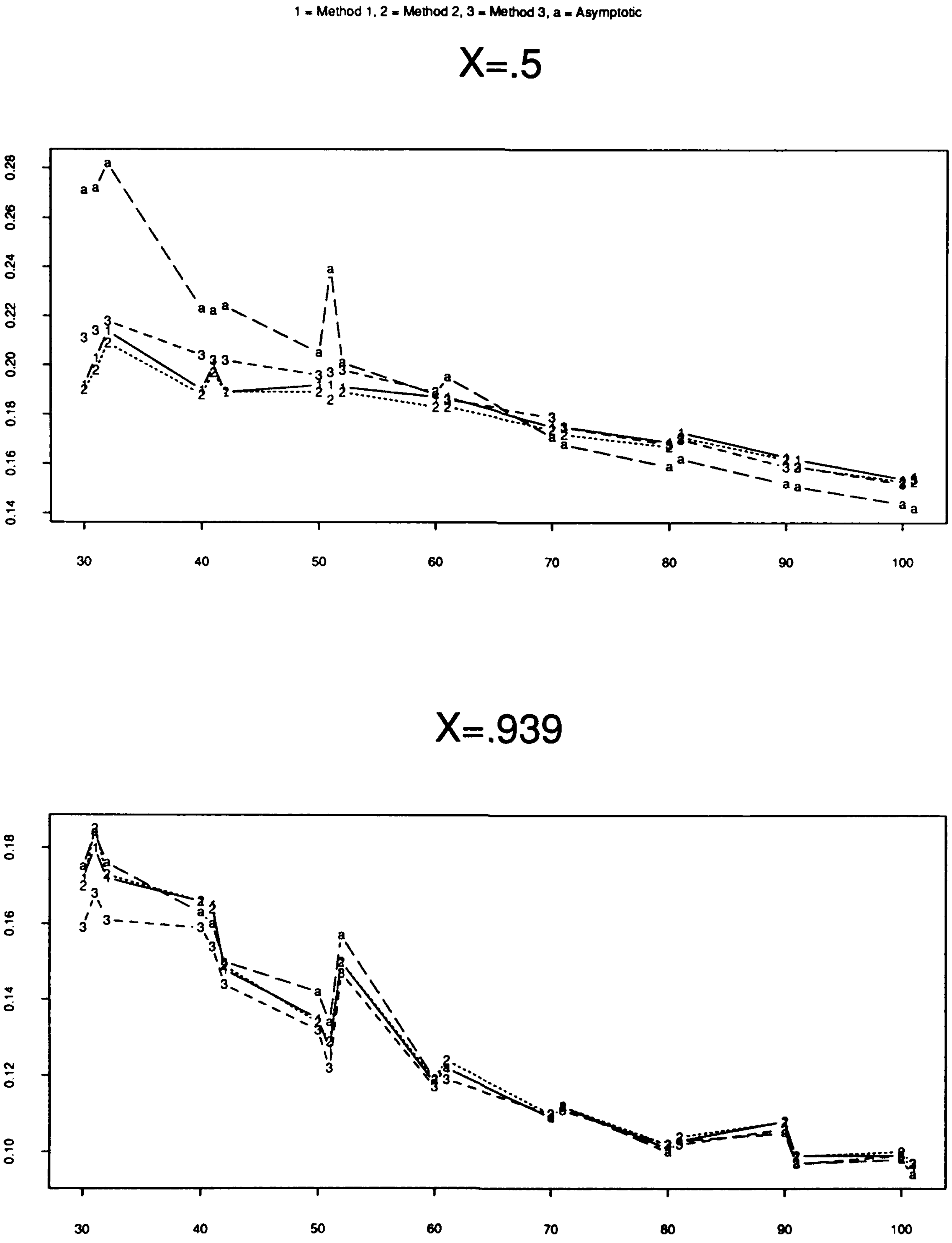


Figure 9: Median Length of Percentile and Hybrid Intervals vs. Sample Size for Median Survival at Two X Values



X=.5

X=.939

Table 1. The Study with an Influential Point: Coverage Probabilities and Average or Median Length of Confidence Intervals.

	β		$S(.10)$		$S(.25)$		$\xi(.54)$		$\xi(.99)$	
	Cov. Pr.	Ave. Len.	Cov. Pr.	Ave. Len.	Cov. Pr.	Ave. Len.	Cov. Pr.	Med. Len.	Cov. Pr.	Med. Len.
P1	.88	2.80	.88	.253	.89	.342	.89	.262	.89	.145
2	.85	2.32	.88	.250	.89	.327	.89	.256	.86	.139
3	.85	2.28	.90	.262	.90	.330	.90	.251	.87	.139
H1	.88	2.80	.84	.253	.89	.342	.77	.262	.78	.145
2	.83	2.32	.84	.250	.87	.327	.76	.256	.77	.139
3	.80	2.28	.84	.262	.86	.330	.75	.251	.76	.139
T1	.81	2.29	.95	.301	.97	.407				
2	.79	2.25	.95	.293	.96	.392				
3	.78	2.24	.95	.293	.95	.381				
A	.84	2.35	.86	.246	.85	.298	.86	.232	.92	.150

Note: P = percentile intervals, H = hybrid intervals, T = boot-*t* intervals, A = asymptotic intervals. 1, 2, 3 refer to Methods 1, 2, and 3, respectively, of forming the bootstrap samples.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 455	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Study of Bootstrap Confidence Intervals in A Cox Model		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s)		8. CONTRACT OR GRANT NUMBER(s) N0025-92-J-1264
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR-042-267
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 111		12. REPORT DATE July 17, 1992
		13. NUMBER OF PAGES 41
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Key words and phrases: ancillarity principle, bootstrap-t, hybrid interval, percentile interval, proportional hazards model

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

See Reverse Side

Abstract

We study bootstrap confidence intervals for three types of parameters in Cox's proportional hazards model: the regression parameter, the survival function at fixed time points, and the median survival time at fixed values of a covariate. Several types of bootstrap confidence intervals are studied, and the type of interval is determined by two factors. One factor is the method of drawing the bootstrap sample. We consider three such methods, which may be briefly described as follows: (1) Ordinary resampling from the empirical cumulative distribution function, (2) Resampling conditional on the covariates, and (3) Resampling conditional on the covariates and the censoring pattern. Another factor is the method of forming the confidence interval from a given sample; the methods considered are the percentile, hybrid, and bootstrap-t. We provide a theorem on the asymptotic validity of the third method of bootstrap resampling. All the methods of forming confidence intervals are compared to each other and to the standard asymptotic method via a Monte Carlo study. The data sets for this Monte Carlo study are simulated conditionally on the covariates and the censoring pattern, the situation appropriate for the third method of resampling. One conclusion drawn from the Monte Carlo study is that the asymptotic method is best for the regression parameter, but not for the survival function or the median survival time. Conclusions about the bootstrap methods include the surprising result that, overall, the second method of drawing the samples outperforms the third method. Also, there is an interaction effect between the two factors, method of drawing the sample and method of forming the interval, especially for estimation of the regression parameter. Finally, the bootstrap-t intervals are consistently outperformed by at least one of the two more rudimentary types of bootstrap interval.